R E P O R T R E S U M E S

ED 019 118 PS 000 826

FINAL REPORT ON HEAD START EVALUATION AND RESEARCH--1966-67 TO THE INSTITUTE FOR EDUCATIONAL DEVELOPMENT. SECTION II, ON THE INTERPRETATION OF MULTIVARIATE SYSTEMS.

BY- LAND, KENNETH C.
TEXAS UNIV., AUSTIN,CHILD DEVELOP.EVAL.AND RES.CTR

REPORT NUMBER IED-66-1 PUB DATE 31 AUG 67
EDRS PRICE MF-$0.50 HC-$3.36 82P.

DESCRIPTORS- *RESEARCH METHODOLOGY, *MATHEMATICAL MODELS, RESEARCH TOOLS, MATHEMATICAL APPLICATIONS, *CRITICAL PATH METHOD, *STATISTICAL ANALYSIS, ANALYSIS OF VARIANCE, CORRELATION, OPERATIONS RESEARCH, LINEAR PROGRAMING,

THIS REPORT PRESENTS A DISCUSSION OF 2 TECHNIQUES WHICH CAN BE USED TO REPRESENT AND INTERPRET MULTIVARIATE STATISTICAL SYSTEMS WHEN IT IS FELT THAT THERE ARE CAUSAL RELATIONS BETWEEN SOME OF THE VARIABLES. THE BASIC TECHNIQUE IS PATH ANALYSIS AND THE OTHER IS ITS EXTENSION THROUGH THE USE OF RECURSIVE SYSTEMS OF EQUATIONS. THE ANALYSIS IS RESTRICTED IN APPLICATION TO RELATIONSHIPS BETWEEN INTERVAL-MEASURABLE VARIABLES THAT ARE LINEAR, ADDITIVE, AND ASYMMETRIC. TO MAKE A PATH ANALYSIS, THE VARIABLES IN THE SYSTEM ARE CLASSIFIED AS EITHER EXOGENOUS, THAT IS, HAVING THEIR VALUES DETERMINED BY FACTORS OUTSIDE THE SYSTEM, OR ENDOGENOUS, THAT IS, HAVING THEIR VALUES DETERMINED BY FACTORS REPRESENTED BY VARIABLES WITHIN THE SYSTEM. BASED ON THIS ANALYSIS, A SET OF REGRESSION EQUATIONS REPRESENTING THESE RELATIONS IS FORMED. THIS SET IS TERMED THE PATH MODEL, AND GRAPHIC CONVENTIONS ARE GIVEN FOR DIAGRAMING IT. THE COEFFICIENTS IN THE EQUATIONS ARE SIMILAR TO THE CORRELATION COEFFICIENTS OCCURRING IN ORDINARY LEAST-SQUARES REGRESSION EQUATIONS. THE ADVANTAGE OF THE PATH ANALYSIS APPROACH IS THAT IT ENABLES THE EXPERIMENTER TO UTILIZE ALL THE INFORMATION AT HIS DISPOSAL, PARTICULARLY THAT CONCERNING CAUSAL RELATIONS BETWEEN VARIABLES. THE TECHNIQUE IS ILLUSTRATED WITH APPLICATIONS TO BIVARIATE AND MULTIVARIATE SYSTEMS HAVING SINGLE AND MULTIPLE STAGES OF CAUSAL INFLUENCE. SOME EXAMPLES DRAWN FROM ACTUAL RESEARCH PROJECTS ARE INCLUDED. (DR)

ED019118

FINAL REPORT ON

HEAD START EVALUATION AND RESEARCH: 1966-67

(Contract No. 66-1)

TO

THE INSTITUTE FOR EDUCATIONAL DEVELOPMENT

By

The Staff and Study Directors

CHILD DEVELOPMENT EVALUATION AND RESEARCH CENTER

John Pierce-Jones, Ph.D., Director

The University of Texas at Austin

August 31, 1967

PS 000825
PS 000826

Section II: ON THE INTERPRETATION OF MULTIVARIATE SYSTEMS

by

Kenneth C. Land

U. S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

FINAL REPORT ON

HEAD START EVALUATION AND RESEARCH: 1966-67

(Contract No. 66-1)

TO

THE INSTITUTE FOR EDUCATIONAL DEVELOPMENT

By

The Staff and Study Directors

CHILD DEVELOPMENT EVALUATION AND RESEARCH CENTER

John Pierce-Jones, Ph.D., Director

The University of Texas at Austin

August 31, 1967

PS 000826

Section II: ON THE INTERPRETATION OF MULTIVARIATE SYSTEMS

by

Kenneth C. Land

PREFACE

This essay is written for research workers in the behavioral sciences. My assumptions regarding this intended audience have several implications for the characteristics of the paper. First, it is not assumed that research workers have great proficiency in the logical manipulation of symbols. Hence, this is no place for advanced mathematical and statistical niceties - rigorous, general, and aesthetically pleasing though they may be. Furthermore, derivations are carried out in great detail and accompanied by extensive exposition. The aim of the paper, although it may be contradictory, is to develop an informal, intuitive rationale of the material for the reader which parallels the formal, rigorous reasoning behind the topics. If this goal is attained, then the researcher should be able to confidently apply the methods to his own empirical problems. Finally, this paper is symbolic of my faith that, for at least certain areas of behavioral science inquiry, the relevant question is no longer "What variables are important?" but "How are the important variables related?" It is my belief that the methods presented in this paper are appropriate to the latter question. On the other hand, straightforward application of statistical principles of estimation and tests of significance are probably more relevant to the former.

It is with great pleasure that I acknowledge my indebtedness to Drs. John Pierce-Jones and Grover Cunningham of the Child Development Evaluation-Research Center for the time to do this research. Because

of the pressure to produce "significant" empirical findings experienced in many behavioral science research centers, the work directive to "be creative" methodologically is all too infrequent. Although I make no claim to origination of any of the notions in this paper, their synthesis herein from diverse sources is my response to the above-mentioned stimulus. This implies, of course, that I am responsible for any errors in the presentation.

Kenneth C. Land

TABLE OF CONTENTS

SECTION		PAGE
1.	Path Analysis	2
1.1.	Path Models and Path Diagrams	3
1.2.	Path Coefficients and Path Regressions	6
1.3.	The Bivariate Path Model	8
1.4.	The Multivariate Path Model	11
1.5.	An Empirical Example of the Multivariate Path Model	22
2.	Recursive Sets of Simultaneous Equations	27
3.	Path Analysis Revisited	31
3.1.	The Multi-stage, Multivariate Path Model	31
3.2.	The Multi-stage, Bivariate Path Model	39
3.3.	The Path Decomposition Model	45
3.4.	The Basic Assumptions of Path Analysis and Model Testing	50
ADDITIONAL READING		54
APPENDIX I		A1
APPENDIX II		A10
BIBLIOGRAPHY		B1

LIST OF TABLES

TABLE		PAGE
1.	Estimators of Path Coefficients for the Bivariate Path Model	10
2.	Estimators of Path Coefficients for the Multivariate Path Model	12
3.	Estimators of Path Coefficients for the Path Model of Figure 7(a)	42
4.	Estimators of Path Coefficients for the Path Model of Figure 7(b)	43
5.	Correlation Matrix for Logarithms of Variables in Duncan's Path Decomposition Model	47

LIST OF FIGURES

FIGURE	PAGE
1	6
2	10
3	12
4	21
5	23
6	34
7	41
8	48

ON THE INTERPRETATION OF MULTIVARIATE SYSTEMS

Kenneth C. Land

This paper constitutes a systemmatic introduction to two procedures which have been developed to aid the representation and interpretation of multivariate statistical systems - path analysis and recursive systems of equations. The first section defines path models and path diagrams. It also develops two elementary applications of the procedure. In the second section, the representation of statistical systems by recursive sets of equations is discussed. Finally, the third section of the paper builds on notions of the two preceding sections by extending path analysis to highly complex systems of relations.

The author has attempted to provide both a systemmatic and, to some extent, complete discussion of the topics. Therefore, two appendices review the basic mathematics of least-squares correlation and regression. If the reader has difficulty understanding the main body of the paper because he has forgotten some basic statistical notions, he may read the appendices and then return to the main sections of the essay. Furthermore, the author has attempted to show that the notions of path analysis, at least for the systems discussed in this paper, follow directly from the basic statistical notions of least-squares correlation and regression. Finally, a main goal of the paper is to develop enough basic understanding on the part of the reader

that he may proceed to utilize the method in his own area of research. Therefore, although the discussion and derivations are on an elementary level, they are accompanied by detailed exposition to provide at least an intuitive and informal understanding of what is really going on.

Section 1: Path Analysis.

The method of path analysis or path coefficients was developed by the geneticist Sewall Wright in a series of general essays (Wright, 1921, 1934, 1954, 1960a, 1960b) as an aid to the quantitative development of genetics. Wright stated the primary purpose of the method in his first general account (1921) as follows:

> The present paper is an attempt to present a method of measuring the direct influence along each *separate path in such a system and thus of find*ing the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degree of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain, the method can be used to find the logical consequences of any particular hypothesis in regard to them.

Wright elaborated the purpose of the method in subsequent papers:

> ... the method of path coefficients is not intended to accomplish the impossible task of deducing causal relations from the values of the correlation coefficients. (1934) ... Path analysis is an extension of the usual verbal interpretation of statistics not of the statistics themselves. It is usually easy to give a plausible interpretation of any significant statistic taken by itself. The purpose of path analysis is to determine whether a proposed set of interpretations is consistent throughout. (1960b)

So much for the intentions of the method. Let us begin by developing the basic notions of path analysis. From that point, we shall develop a few simple, almost trivial, applications of the path notions. Finally, we shall, after the introduction of some additional notions in the next section of the paper, proceed to the path analysis of complex systems of relations on variables in which the advantages of the procedure begin to accumulate in such a manner as to make path analysis worth the effort of becoming proficient in the method.

1.1. Path Models and Path Diagrams. We begin by restricting the application of the method to sets of relationships among variables which are (1) linear, (2) additive, and (3) asymmetric. Furthermore, the variables must be measurable or be conceived as measurable on an interval scale, although some of them may not actually be measured. We shall return to these assumptions at the end of the paper.

In such systems of relationships, a subset of the variables is taken as linearly dependent on the remaining variables, which are assumed to be independent. That is, the total variation of the independent variables is assumed to be caused by variables outside of the set under consideration. We may refer to such variables as "exogenous." The exogenous variables in a particular set may be correlated among themselves; however, the explanation of their intercorrelation is not a problem for the system under consideration. The subset of variables which are taken as dependent variables in the total set may be termed "endogenous" variables. In contrast to the exogenous subset of variables, the total variation of the endogenous variables is assumed to

be completely determined by some combination of the variables in the system. Note that this implies that, in some path models, a subset of the endogenous variables may be conceived to have causal effects on other endogenous variables in addition to the direct effects of the exogenous variables. Furthermore, in those systems of relationships where an endogenous variable is not completely determined by prior (exogenous or endogenous) measured variables, a residual variable uncorrelated with the set of variables immediately determining the variable under consideration is introduced to account for the variance of the dependent variable not explained by measured variables. The basic assumptions of path analysis have been reviewed in these two paragraphs. Because they are so basic, the reader may find it useful to re-read the assumptions several times as the method is developed below.

The notion of the path diagram was developed by Wright (1921, 1934, 1960a) to provide a convenient representation of those systems of relations which conform to the assumptions of the above paragraphs. Path diagrams are drawn according to the following conventions:

(1) The postulated causal relations among the variables of the system are represented by uni-directional arrows extending from each determining variable to each variable dependent on it.

(2) The postulated non-causal correlations between exogenous variables of the system are symbolized by two-headed curvilinear arrows to distinguish them from causal arrows.

(3) Residual variables are also represented by uni-directional

arrows leading from the residual variable to the dependent variable. However, literal subscripts are attached to residual symbols to indicate that these variables are not measured.

(4) Finally, the quantities entered beside the arrows on a path diagram are the symbolic or numerical values of the path and correlation coefficients of the postulated relationships. The symbolic form of the path coefficient is P_{ij}, where the first subscript i denotes the dependent variable and the second subscript j denotes the variable whose determining influence is under consideration. Note that, since we are considering only asymmetric causal relations, the coefficients P_{ij} and P_{ji} will never appear in the same path diagram together, i.e., either P_{ij} or P_{ji} but never both will be postulated in a given system. Furthermore, the coefficient P_{ij} will ordinarily be a partial path coefficient; however, we do not denote the variables held constant after a dot as with ordinary least-squares partial regression and correlation coefficients. They will usually be obvious from the path diagram.

In this paper, we shall use the term path model to refer to the regression equation or set of regression equations which represents the postulated causal and non-causal relationships among the variables under consideration. A property of path diagrams which conform to the above rules of representation is an isomorphism with the algebraic and statistical properties of the postulated system of relationships. In other words, there is a one-to-one correspondence between the postulated causal and non-causal relations of a path model

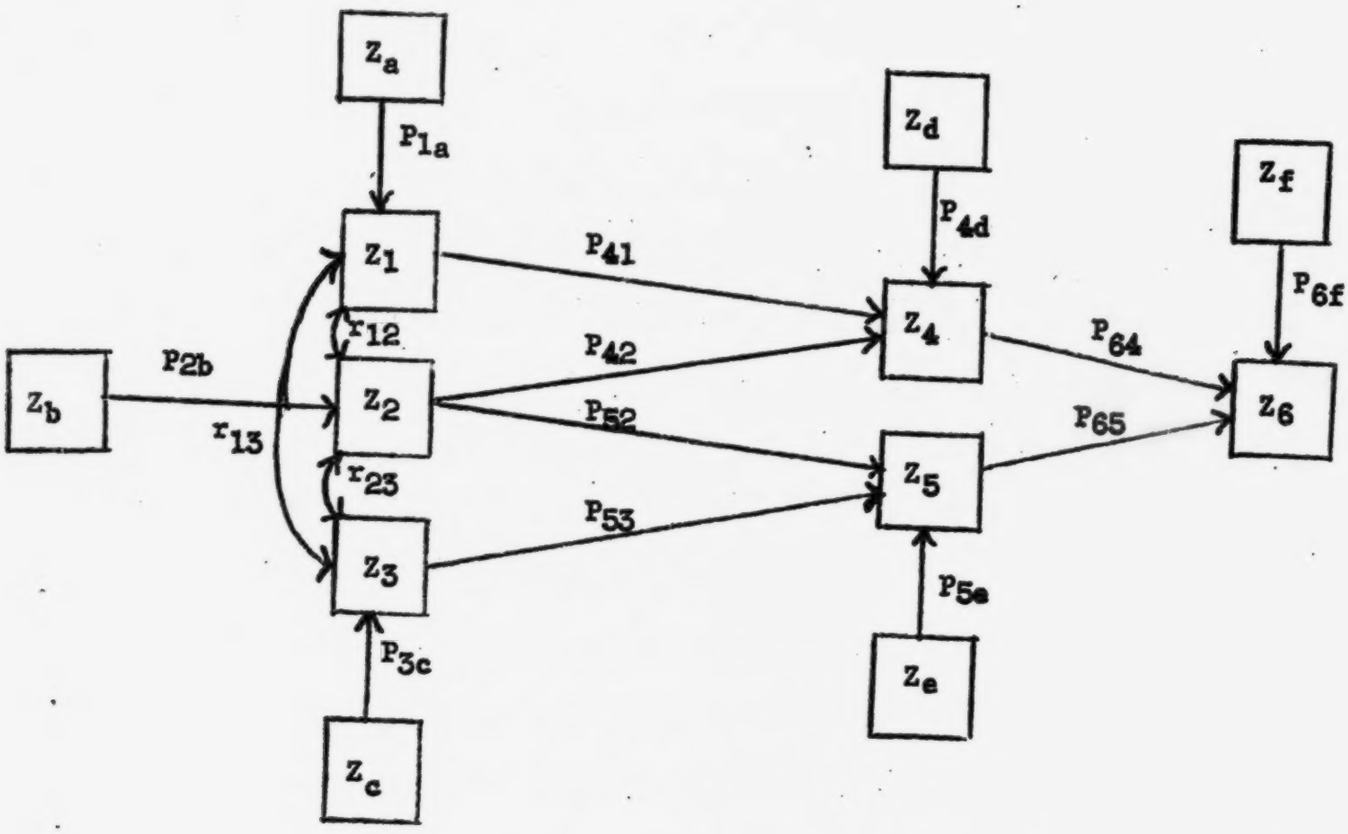


FIGURE 1.

and its path diagram. This property and its usefulness will become more obvious as we develop the method. As an illustration of the conventions of path diagrams, we have a possible system in Figure 1.

1.2. Path Coefficients and Path Regressions. Since all relations are assumed to be linear, we may write the dependent relationship of X_1 on $X_2, X_3, \ldots, X_n$, and residual X_a, in raw-score form as follows:

$$X_1 = c_{12}X_2 + c_{13}X_3 + \ldots + c_{1n}X_n + c_{1a}X_a \qquad (1.1)$$

or, in deviation-units, we have

$$(X_1-M_1) = c_{12}(X_2-M_2) + c_{13}(X_3-M_3) + \ldots + c_{1n}(X_n-M_n) + c_{1a}(X_a-M_a)$$

where M_i is the mean of the ith variable. Letting $x_i = (X_i-M_i)$, this is

$$x_1 = c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n + c_{1a}x_a \qquad (1.2)$$

It is often more convenient to utilize each variable in standard-unit form. Let $Z_i = (X_i - M_i)/S_i$ and $P_{1i} = c_{1i}(S_i/S_1)$, where S_i denotes the standard-deviation of the ith variable. Then formula (1.2) becomes

$$Z_1 = P_{12}Z_2 + P_{13}Z_3 + \ldots + P_{1n}Z_n + P_{1a}Z_a \qquad (1.3)$$

The coefficients c_{12}, etc., where the first subscript denotes the dependent variable while the second subscript denotes the independent variable, are of the type of partial regression coefficients but may exist in a system with unmeasured hypothetical variables in addition to the residual variable. Hence, we shall refer to them as <u>path regression coefficients</u>. Also, the standardized coefficients P_{12}, etc., are of the type called <u>path coefficients or standardized path coefficients</u>. <u>Each path coefficient measures</u> the fraction of the standard deviation of the dependent variable (with the appropriate sign) for which the designated variable is <u>directly</u> responsible in the sense of the fraction which would be found if this factor varies to the same extent as in the observed data while all other variables (including residual factors) are constant (Wright, 1934). This definition (except for determination of sign) can be written as follows:

$$P_{12} = \frac{S_{1\cdot 23\ldots n,a}}{S_1} \cdot \frac{S_2}{S_{2\cdot 34\ldots n,a}}$$

$$= \frac{S_2}{S_1} \cdot \frac{S_{1\cdot 23\ldots n,a}}{S_{2\cdot 34\ldots n,a}}$$

$$= \frac{S_2}{S_1} \cdot c_{12} \qquad (1.4)$$

where $S_{i \cdot jk \ldots n,a}$ indicates the standard deviation of the ith variable with variables j through n and residual variable a held constant. Given this definition of the path coefficient, it is obvious that the squared path coefficient measures the portion of the variance of the dependent variable for which the independent variable is directly responsible.

Now that we have introduced the definition of path coefficients and path regressions, the alert reader will immediately note a similarity of path regressions to ordinary least-squares partial regression coefficients and a similarity of path coefficients to least-squares standardized partial regression coefficients or beta weights as discussed in the two appendices of this paper. It is true that for certain kinds of systems of relationships the path coefficients and regressions are identical to the least-squares estimators. However, this is not true in general (see Wright, 1934, 1954, 1960a). It will be our task to develop the method of path coefficients only for these cases in which path coefficients and regressions are identical to the least-squares estimators for correlation and regression coefficients in the remainder of this paper. We shall find that path analysis yields information about a statistical system which helps render an interpretation possible. It does so, not by additional statistical analysis, but primarily by forcing the researcher to utilize all of the information at his disposal. We proceed to develop the procedure for elementary applications.

1.3. The Bivariate Path Model. The simplest type of relation to which path analysis may be applied is the case of a dependent variable

X_2, independent or exogenous variable X_1, and residual variables X_a and X_b:

$$x_2 = c_{21}x_1 + c_{2b}x_b$$

or, in standard-form, the path model is

$$z_2 = P_{21}z_1 + P_{2b}z_b \qquad (1.5)$$

This is simply a case of bivariate least-squares regression with explicit consideration of the residual term. Figure 2 is a path diagram for this model. Note that, since Z_1 is considered exogenous, $P_{1a} = 1.0$, i.e., the total variation of Z_1 is caused by unmeasured variables or variables outside the present model. Because this is true for all exogenous variables, symbols for their residuals and residual path coefficients are often dropped from the diagram in the interest of neatness of representation.

In this model, as shown in Table 1, the least-squares estimator for the path cefficient P_{21} is the correlation coefficient r_{21} which is also equal to the beta weight in the bivariate case as is shown in the first appendix of this paper (section A.1.1.). Here, as in the appendices, we symbolize least-squares standardized partial regression coefficients or beta weights by B^*_{ij} since we do not have a Greek letter for beta on our keyboard. Now, let us consider the estimate and meaning of the residual path coefficient. Since Z_b is independent of Z_1 (it was stated earlier that the residual is independent of the immediate predictors in a model and this also follows from least-squares estimation principles), we know by the same principle as for Z_1 that $r_{2b} = B^*_{2b} = P_{2b}$. However, since the residual represents all variables outside the system which cause variation in Z_2, it is unmeasured and we do not have a direct estimate for P_{2b}.

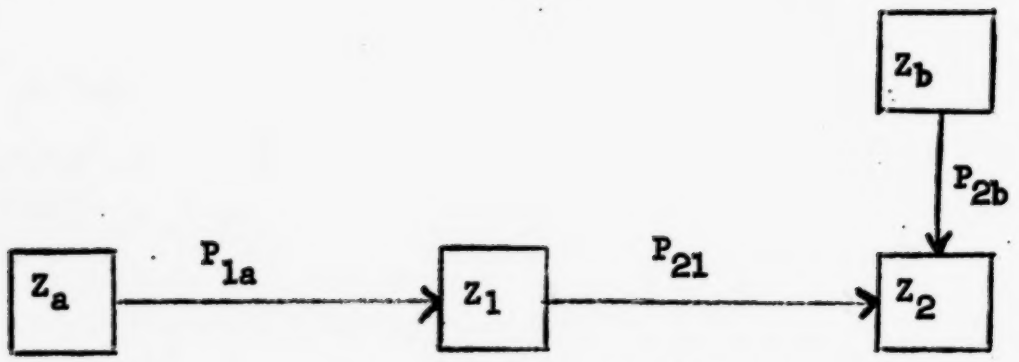


FIGURE 2.

TABLE 1: ESTIMATORS OF PATH COEFFICIENTS FOR THE BIVARIATE PATH MODEL

(1) Path Model: $Z_2 = P_{21}Z_1 + P_{2b}Z_b$

(2) $P_{21} = r_{21} = B^*_{21}$

(3) $r_{22} = 1 = P^2_{21} + P^2_{2a}$

(4) $P_{2a} = \sqrt{1 - P^2_{21}} = \sqrt{1 - r^2_{21}}$

Therefore, we must estimate it indirectly by utilizing the fact that the squares of P_{21} and P_{2b} must sum to unity. That is, since the square of each of P_{21} and P_{2b} is the proportion of the variance of Z_2 explained by Z_1 and Z_b, respectively, and the variables are independent, the sum of the proportions must be unity. Hence, P_{2b} is the square root of the quantity - one minus P^2_{21} or one minus r^2_{21} - as is shown in Table 1. This looks familiar. It is in fact what we define as the coefficient of alienation in the first appendix of the paper (section A.1.2.). Now we see the first contribution of path analysis to the interpretation of regression systems: It provides a convenient and logically sound interpretation of the coefficient of alienation as the path coefficient of the residual term in the regression equation. It may be shown that the residual variable

has a zero mean and unit variance or standard deviation since we have conceived of all of the variables as being in standard-form. Hence, it may be helpful for the reader to think of the residual as a dummy variable having unit variance and zero mean and representing all unmeasured variables which cause variation in the dependent variable. The residual path coefficient, then, is the proportion of the standard deviation, and its square is the proportion of the variance, of the dependent variable which is caused by all (unmeasured) variables outside of the set under consideration in the path model. It should be noted that this interpretation of the residual path coefficient is consistent with the derivations in Appendix I of the paper (equations A1.9, A1.10, and A1.11). However, the hard-nosed reader who does not trust the above logic may easily demonstrate the relation in a specific empirical case. If he has a computer regression program available, he may simply predict a criterion variable from a predictor variable, print out the residual variable, and then predict the criterion from the residual. The beta weight in the second prediction should be comparable to the square root of one minus r-square in the first prediction.

1.4. The Multivariate Path Model. We shall now extend path analysis to the multivariate case. Consider a path model in which an endogenous variable Z_3 is dependent upon exogenous variables Z_1 and Z_2 and a residual variable Z_a:

$$Z_3 = P_{31}Z_1 + P_{32}Z_2 + P_{3a}Z_a$$

A path diagram of this model is given in Figure 3, and the estimators

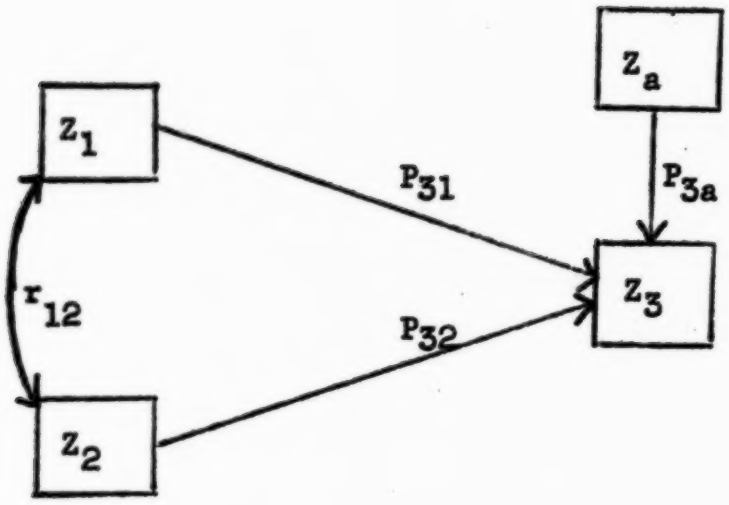


FIGURE 3.

TABLE 2: ESTIMATORS OF PATH COEFFICIENTS FOR THE MULTIVARIATE PATH MODEL OF FIGURE 3.

Standardized Path Coefficients

(1) Path Model for Figure 3: $Z_3 = P_{31}Z_1 + P_{32}Z_2 + P_{3a}Z_a$

(2) $r_{31} = P_{31} + P_{32}\, r_{12}$

(3) $r_{32} = P_{31}\, r_{12} + P_{32}$

(4) r_{12} given

(5) $r_{33} = 1 = P_{31}r_{31} + P_{32}r_{32} + P_{3a}^2$

(6) $P_{3a}^2 = 1 - P_{31}r_{31} + P_{32}r_{32} = 1 - R^2$

(7) $P_{3a} = \sqrt{1 - R^2}$

Path Regression Coefficients

(8) $b_{31} = c_{31} + c_{32}b_{21}$

(9) $b_{32} = c_{31}b_{12} + c_{32}$

(10) b_{21} and b_{12} given

for the path coefficients and path regressions are shown in Table 2.

There are a number of important principles to be gained from an analysis of Figure 3 and Table 2. First, let us derive the relation of the standardized path coefficients P_{ij} and the path regression coefficients c_{ij} to the least-squares correlation coefficients r_{ij} and regression coefficients b_{ij}. For example, take the case of the path coefficient P_{31}. We are given

$$r_{31} = P_{31} + P_{32}r_{12} \tag{1.6}$$

so, by substracting $P_{32}r_{12}$ from both sides we have

$$P_{31} = r_{31} - P_{32}r_{12}$$

By the same principle, $P_{32} = r_{32} - P_{31}r_{12}$. So, by substitution

$$\begin{aligned} P_{31} &= r_{31} - (r_{32} - P_{31}r_{12})\, r_{12} \\ &= r_{31} - r_{32}r_{12} + P_{31}r_{12}r_{12} \end{aligned}$$

Then, by subtracting $P_{31}r_{12}^2$ from both sides of the equation, the equation becomes

$$P_{31} - P_{31}r_{12}^2 = r_{31} - r_{32}r_{12}$$

or, by factoring out P_{31} on the left-hand side,

$$P_{31}\,(1 - r_{12}^2) = r_{31} - r_{32}r_{12}$$

whence, by dividing both sides by $(1 - r_{12}^2)$, we have

$$P_{31} = \frac{r_{31} - r_{32}r_{12}}{1 - r_{12}^2} \tag{1.7}$$

In this form, P_{31} bears immediate similarity to formula (A2.2) and is, in fact, identical with the least-squares estimator for the standardized partial regression coefficient. We may derive a similar formula for the path

regression coefficients c_{ij}. The derivation for c_{31} is as follows:

$$b_{31} = c_{31} + c_{32}b_{21} \tag{1.8}$$

$$\begin{aligned} c_{31} &= b_{31} - c_{32}b_{21} \\ &= b_{31} - (b_{32} - c_{31}b_{12})\, b_{21} \\ &= b_{31} - b_{32}b_{21} + c_{31}b_{12}b_{21} \\ c_{31} - c_{31}b_{12}b_{21} &= b_{31} - b_{32}b_{21} \\ c_{31}(1 - b_{12}b_{21}) &= b_{31} - b_{32}b_{21} \end{aligned}$$

So,

$$c_{31} = \frac{b_{31} - b_{32}b_{21}}{1 - b_{12}b_{21}} \tag{1.9}$$

Again, in this form, c_{31} is identical to the least-squares estimator for partial regression coefficient in raw- or deviation-score form as given in Appendix formula (A2.5). The above two derivations demonstrate that <u>for the multivariate path model</u>, of which Figure 3 is a special case, <u>the path coefficients and path regressions are equivalent to the least-squares estimators</u> for the standardized and raw- or deviation-score partial regression coefficients, respectively.

<u>A second principle</u> that emerges from Figure 3 and Table 2 is illustrated by formulas (2) and (3) of Table 2. We know that we are given the correlations of the exogenous variables in any path model (r_{12} in Figure 3) and we are not interested in them. But formulas (2) and (3) show the basic composition of the correlation of each exogenous variable with the endogenous variable. Let us explore the derivation of, say, formula (3). By definition of the correlation coefficient for variables

in standard-score form, we have

$$r_{31} = (1/N) \Sigma Z_3 Z_1 \qquad (1.10)$$

where the summation is over all observed Z-values for variables 1 and 3. The convention regarding summations followed in this paper is that, if a summation index is not provided, the summation is over observed values; otherwise, a summation index and explanatory note will be provided. Regarding (1.10), we know that Z_3 - the standard-score form of the criterion variable - is totally dependent upon Z_1, Z_2, and Z_a. Hence, by substitution from formula(1) of Table 2 - the path model, we have

$$r_{31} = (1/N) \Sigma Z_1 (P_{31}Z_1 + P_{32}Z_2 + P_{3a}Z_a) \qquad (1.11)$$

or, by expansion,

$$= (1/N) (P_{31} \Sigma Z_1 Z_1 + P_{32} \Sigma Z_1 Z_2 + P_{3a} \Sigma Z_1 Z_a)$$

$$= P_{31} \frac{\Sigma Z_1^2}{N} + P_{32} \frac{\Sigma Z_1 Z_2}{N} + P_{3a} \frac{\Sigma Z_1 Z_a}{N} \qquad (1.12)$$

and since (1) the sum of the squares of standard-scores for a variable is unity, (2) the sum of cross-products of standard-scores for two variables is the correlation coefficient of the variables, and (3) the correlation of the residual Z_a with an immediate determining variable of Z_3 is zero, (1.12) reduces to

$$r_{31} = P_{31} + r_{12}P_{32} \qquad (1.13)$$

and we have derived equation (2) of Table 2.

There are several insights to be derived from (1.13). First, it implies that, for a multivariate path model, the correlation of an

exogenous variable and the dependent variable is the sum of the direct effect via its path coefficient from that exogenous variable to the dependent variable and its indirect effect(s) through its correlation with the other exogenous variable(s) as measured by the product of the correlation coefficient of the two exogenous variables and the path coefficient of the latter variable. This is a second contribution of path analysis to the interpretation of regression systems: It provides an interpretation of the correlation of a predictor and the criterion as a sum of direct and indirect effects. This interpretation is certainly not obvious from the typical formulasfor the correlation and beta coefficients! Of course, a similar type of interpretation holds for formulas (8) and (9) of Table 2 regarding total and partial regression coefficients in raw- or deviation-score form.

Let us look again at formula (1.13). If the total correlation between the exogenous variable Z_1 and the endogenous variable Z_3 is made up of a sum of direct and indirect effects, and if the direct effect is estimated by P_{31}, then the indirect effect must be estimated by $r_{12}P_{32}$, or in a more generally applicable form:

$$\text{Total Indirect Effect (TIE) of } Z_1 \text{ on } Z_3 = r_{31} - P_{31} \qquad (1.14)$$

Thus, we have an example of the third contribution of path analysis to the interpretation of regression systems: It provides a general procedure for exploring the indirect effects of an independent variable on a dependent variable in a multivariate path model. This contribution will become more interesting as the path models become more complex.

Equation (5) of Table 2 may be viewed as a special case of (2) and (3) - the case of complete determination of the dependent variable. Let us explore the derivation of this equation. We know, as before that the formula for the correlation of Z_3 with itself is:

$$r_{33} = 1 = (1/N) \Sigma Z_3 Z_3 \qquad (1.15)$$

By substitution of formula (1) of Table 2, this becomes

$$\begin{aligned} r_{33} = 1 &= (1/N) \Sigma Z_3 (P_{31}Z_1 + P_{32}Z_2 + P_{3a}Z_a) \\ &= (1/N) (P_{31} \Sigma Z_3 Z_1 + P_{32} \Sigma Z_3 Z_2 + P_{3a} Z_3 Z_a) \\ &= P_{31} \frac{\Sigma Z_3 Z_1}{N} + P_{32} \frac{\Sigma Z_3 Z_2}{N} + P_{3a} \frac{\Sigma Z_3 Z_a}{N} \end{aligned}$$

For the same reasons as given above for the derivation of (1.13) and the fact that, since Z_a is independent of Z_1 and Z_2, $r_{3a} = P_{3a}$, this becomes

$$\begin{aligned} r_{33} = 1 &= P_{31}r_{31} + P_{32}r_{32} + P_{3a}^2 \qquad (1.16) \\ &= \sum_{i=1}^{2} P_{3i}r_{3i} + P_{3a}^2 \end{aligned}$$

which is (5) of Table 2. This equation may be further expanded by the substitution of values for r_{31} and r_{32} as follows

$$\begin{aligned} r_{33} = 1 &= P_{31}(P_{31}+P_{32}r_{12}) + P_{32}(P_{31}r_{12}+P_{32}) + P_{3a}^2 \\ &= P_{31}^2 + P_{31}P_{32}r_{12} + P_{32}P_{31}r_{12} + P_{32}^2 + P_{3a}^2 \\ &= \sum_{i=1}^{2} P_{3i}^2 + 2 \sum_{k,j=1}^{2} P_{3j}P_{3k}r_{jk} + P_{3a}^2 \quad (k>j) \qquad (1.17) \end{aligned}$$

where the range of j and k, (k>j), includes all measured variables.

Let us examine the implications of equations (1.16) and (1.17). First, note that the single summation in (1.16) is equal to the sum of the two summations in (1.17). Second, note that (1.17) is identical to equation (A2.13) of Appendix Two except for the explicit consideration of the residual path coefficient squared - P_{3a}^2. This implies that the sum of the two summations in (1.17), and, therefore, the single summation in (1.16), is equal to R^2 - the square of the multiple correlation coefficient for the path model. This means that we can derive a simple formula for the computation of the residual path coefficient as follows:

$$P_{3a}^2 = 1 - \sum_{i=1}^{2} P_{3i}r_{3i}$$

$$= 1 - \sum_{i=1}^{2} P_{3i}^2 + 2 \sum_{k,j=1}^{2} P_{3j}P_{3k}r_{jk} \quad (k>j)$$

$$= 1 - R^2$$

So,

$$P_{3a} = \sqrt{1 - R^2} \qquad (1.18)$$

A third implication of (1.17) concerns the important problem of the interpretation of the path coefficients in the multivariate path model: Equation (1.17) shows that the total variance of Z_3 is a sum of the squares of each path coefficient plus a term which measures the correlational influence of the exogenous variables. Or, in other words, by multiplying both sides of (1.17) by S_3^2, it may be seen that the squared path coefficients measure the portions of S_3^2 that are determined directly by the exogenous variables while the other summation (which may be negative) measures correlational determination. Now we come to a unique

characteristic of multivariate path coefficients or beta weights as opposed to their bivariate counterparts: Whereas bivariate path or beta coefficients, since they are identical with the correlation coefficient, are bounded by + and - 1.0, the multivariate path coefficients may exceed + 1 or - 1 in absolute value; hence, the square of a multivariate path coefficient may exceed + 1. This would seem to indicate that the independent variable with a path coefficient-squared greater that 1.0 causes more than 100 per cent of the variance in the dependent variable. But this is impossible. So the question is: How does one interpret a squared multivariate path coefficient greater that 1.0? Wright (1960a) gives the following interpretation:

> Such a value shows at a glance that direct action of the factor in question is tending to bring about greater variability than is actually observed. The direct effect must be offset by opposing correlated effects of other factors.

In short, the key to this difficulty would seem to lie in the definition of the path coefficient as given in equation (1.4) and in relation (1.17): If a multivariate path coefficient is greater than + or - 1, then one or the other of the terms in (1.4) must be greater that + or - 1. Furthermore, the correlation of the exogenous variable with the other exogenous variable(s) as in equation (1.17) must be such as to compensate for the tendency of the exogenous variable to cause more variation in the dependent variable than is observed in the data of concern. In a specific empirical example, it may be fruitful to examine each of the specific terms of the second summation in (1.17) to provide insight as to how the correlation of the independent variable of interest with the other exogenous variables is compensating for its

large direct path coefficient. This exploration may lead to insight, for example, as to how the path model could be altered to achieve the same multiple correlation with fewer exogenous variables. As for the residual path coefficient, its interpretation remains the same as for the bivariate path model.

For the reader who is weary of this abstract discussion of multivariate path models, we shall give an empirical example. However, before we leave this topic, let us generalize several important formulas to an n variable multivariate model. The general multivariate path model is

$$Z_1 = P_{12}Z_2 + P_{13}Z_3 + \cdots + P_{1n}Z_n + P_{1a}Z_a \qquad (1.19)$$

where Z_1 is arbitrarily taken as representing the endogenous variable, $Z_2, \ldots, Z_n$ as exogenous variables, and Z_a as residual. The general formula for the correlation of any exogenous variable with the endogenous variable becomes

$$r_{1i} = P_{1i} + \sum_{\substack{j=2 \\ j\neq i}}^{n} P_{1j}\, r_{ij}$$

$$= \sum_{j=2}^{n} P_{1j}\, r_{ij} \qquad (1.20)$$

Then the formula for the total indirect effect of any exogenous variable Z_i on the endogenous variable Z_1 is

$$\text{Total Indirect Effect (TIE) of } Z_i \text{ on } Z_1 = r_{1i} - P_{1i} \qquad (1.21)$$

Also, the formula for the complete determination of Z_1 becomes

$$r_{11} = \sum_{i=1}^{n} P_{1i}^2 + 2 \sum_{k,j=1}^{n} P_{1j} P_{1k} r_{jk} + P_{1a}^2 \qquad (1.22)$$

Finally, the formula for the residual in the general multivariate path model is

$$P_{1a} = \sqrt{1 - R^2} \qquad (1.23)$$

Figure 4 illustrates the general multivariate path diagram.

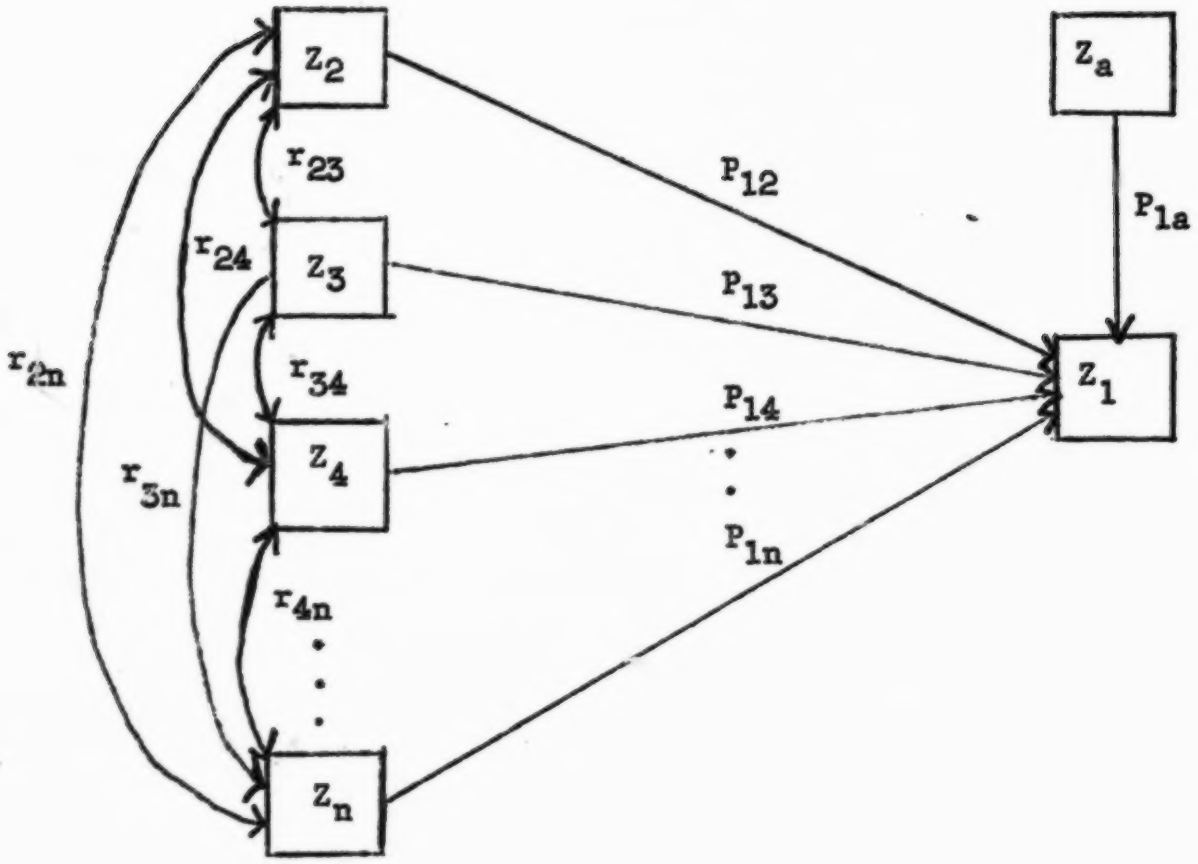


FIGURE 4.

1.5. An Empirical Example of the Multivariate Path Model. This empirical problem involves data kindly provided by Dr. Grover Cunningham of the Child Development Evaluation-Research Center. He postulated the dependency of a child's IQ score on 17 measures of correlation of personality characteristics of his parents. The exact form of his path model is

$$\begin{aligned} Z_{18} = {} & P_{18,1}Z_1 + P_{18,2}Z_2 + P_{18,3}Z_3 + P_{18,4}Z_4 + P_{18,5}Z_5 \\ & + P_{18,6}Z_6 + P_{18,7}Z_7 + P_{18,8}Z_8 + P_{18,9}Z_9 \\ & + P_{18,10}Z_{10} + P_{18,11}Z_{11} + P_{18,12}Z_{12} + P_{13,13}Z_{13} \\ & + P_{18,14}Z_{14} + P_{18,15}Z_{15} + P_{18,16}Z_{16} + P_{18,17}Z_{17} \\ & + P_{18,a}Z_a \end{aligned} \tag{1.24}$$

where Z_i, $i = 1,\ldots,17$ are measures of correlation of personality attributes of parents, Z_{18} is the child's IQ score, and Z_a is the residual factor. A path diagram of the postulated model is given in Figure 5. All of the postulated causal relations are drawn in Figure 5. However, note that none of the postulated correlational relationships are drawn in the figure. The reason for this departure from the norm is that, as the number of exogenous variables increases in a multivariate path model, the number of correlational relationships between the exogenous variables increases according to the formula for the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!\ (n-k)!} \tag{1.25}$$

where n is the total number of exogenous variables, k is the number

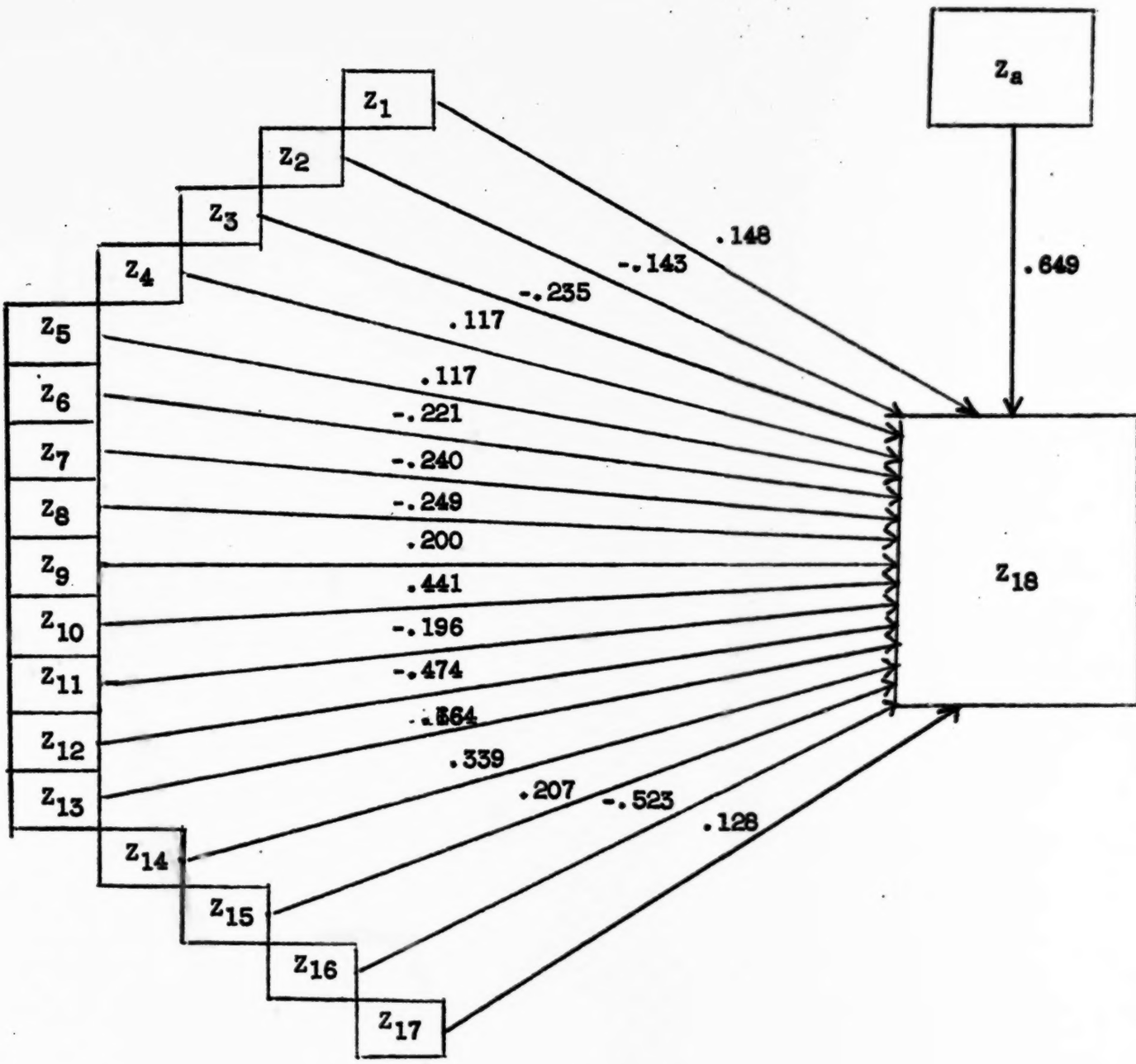

z_a
z_1
z_2
z_3
z_4
z_5
z_6
z_7
z_8
z_9
z_{10}
z_{11}
z_{12}
z_{13}
z_{14}
z_{15}
z_{16}
z_{17}
z_{18}
.148
-.143
-.235
.117
.117
-.221
-.240
-.249
.200
.441
-.196
-.474
-.364
.339
.207
-.523
.128
.649

FIGURE 5.

of variables taken in a combination - this will nearly always be 2 for multivariate path models, and ! is the factorial operator. Hence, the number of correlations between the 17 exogenous variables taken 2 at a time would be:

$$\binom{17}{2} = \frac{17!}{2!\ 15!}$$

$$= \frac{17 \cdot 16 \cdot \cancel{15!}}{2!\ \cancel{15!}}$$

$$= \frac{272}{2} = 136 \quad \text{correlations}$$

In short, the representation of all 136 correlations among the exogenous variables would make Figure 5 look more like abstract art that a path diagram. Therefore, we shall leave correlational arrows out of Figure 5 and choose to keep in mind the posited intercorrelations among the exogenous variables.

The numerical rather than the symbolic values for the path coefficients have been entered in Figure 5. From these values, it is obvious that exogenous variables 10, 12, and 16 make the largest direct contributions to the variance of Z_{18}. However, the residual variable Z_a still accounts for about 42 per cent of the variance of children's IQ scores. The equation for total determination of Z_{18} with the actual numerical values for the direct effects, correlational effects, and residual effects corresponding to equation (1.23) is

$$r_{18,18} = 1.0 = 1.2654 - .6869 + .4215 \qquad (1.26)$$

This equation shows that the correlational effect is negative, quite

large, and counteracts the overdetermination of the IQ scores through the direct effects of the seventeen exogenous variables.

Let us look at the total indirect effects of each of the variables:

TIE of 1 on 18 = $r_{18,1} - P_{18,1}$ = .153 - .148 = .005

TIE of 2 on 18 = $r_{18,2} - P_{18,2}$ = -.089 - (-.143) = .054

TIE of 3 on 18 = $r_{18,3} - P_{18,3}$ = -.021 - (-.235) = .214

TIE of 4 on 18 = $r_{18,4} - P_{18,4}$ = -.011 - .117 = -.128

TIE of 5 on 18 = $r_{18,5} - P_{18,5}$ = .102 - .117 = -.015

TIE of 6 on 18 = $r_{18,6} - P_{18,6}$ = -.227 - (-.221) = -.006

TIE of 7 on 18 = $r_{18,7} - P_{18,7}$ = -.105 - (-.240) = .135

TIE of 8 on 18 = $r_{18,8} - P_{18,8}$ = -.068 - (-.249) = .181

TIE of 9 on 18 = $r_{18,9} - P_{18,9}$ = -.139 - .200 = .061

TIE of 10 on 18 = $r_{18,10} - P_{18,10}$ = -.068 - .441 = -.373

TIE of 11 on 18 = $r_{18,11} - P_{18,11}$ = -.088 - (-.196) = .108

TIE of 12 on 18 = $r_{18,12} - P_{18,12}$ = -.074 - (-.474) = .400

TIE of 13 on 18 = $r_{18,13} - P_{18,13}$ = .120 - .164 = -.044

TIE of 14 on 18 = $r_{18,14} - P_{18,14}$ = .219 - .339 = -.120

TIE of 15 on 18 = $r_{18,15} - P_{18,15}$ = .161 - .207 = -.046

TIE of 16 on 18 = $r_{18,16} - P_{18,16}$ = -.260 - (-.523) = .263

TIE of 17 on 18 = $r_{18,17} - P_{18,17}$ = .093 - .128 = -.035

There are two observations which should be made regarding the estimates of the total indirect effect of each exogenous variable. First, note that, although the variables with large direct effects tend also to

have large indirect effects, the rank-order of the variables by size of indirect effects is quite unlike the rank-order according to size of direct effects. Second, note that, for the three-variable multivariate path model discussed in section 1.4, the total indirect effect was the only indirect effect for each exogenous variable, e.g., variable one had only one indirect effect - through its correlation with variable two and the direct effect of variable two on the endogenous variable. However, in this problem with seventeen exogenous variables, each exogenous variable has sixteen different indirect effects, i.e., each exogenous variable has an indirect effect through its correlation with each of the other exogenous variables. In general, if there are n exogenous variables in a multivariate path model, then there will be n - 1 indirect effects for each exogenous variable. This result follows directly from formula (1.20). As an example, in the present model the indirect effect of variable 1 on variable 18 through variable 2 is the product $r_{12} \cdot P_{18,2} = (-.250)(-.143) = 0.036$ while its indirect effect through variable 3 is the product $r_{13} \cdot P_{18,3} = (-.170)(-.235) = 0.040$. Of course, the sum of all of the indirect effects of an exogenous variable through all of the other exogenous variables is equal to the total indirect effect of the variable. As an aid to the interpretation of a specific empirical problem, it may be useful to examine the separate indirect effects of each variable. It is out of such painstaking empirical examinations that a multivariate behavioral science will be built!

Section 2: Recursive Sets of Simultaneous Equations.

Up to this point in the paper, we have developed path models for elementary multivariate systems. By "elementary" we mean systems of variables such that the postulated relationship is of an endogenous variable dependent upon a number of exogenous variables which are taken to be caused by variables outside of the set under consideration. The reader who has systematically read the preceding section should be able to construct and interpret a multivariate path model for any elementary multivariate system he may encounter. However, the elementary multivariate path model is not sufficient for all of the types of multivariate systems which the behavioral scientists frequently must analyze. Specifically, we often are willing to postulate that an exogenous variable effects an endogenous through its direct effect on another variable. Or, in other situations, we are willing to postulate the causal dependency of what we had considered an exogenous variable. This type of multivariate system is illustrated by Figure 1 of section 1.1. In short, we often wish to isolate "stages" of causation. We shall take a brief excursion from path analysis in this section to develop a tool which will allow us to represent such multivariate systems with path models.

Obviously, the problem posed in the preceding paragraph is the simultaneous representation of several relationships among a set of variables rather than one particular relationship taken by itself. Furthermore, the reader has undoubtedly been exposed to the dictum that in order to represent several relationships at the same time one must write

and solve a set of simultaneous equations. The question is: which of the many possible sets of simultaneous equations are consistent with our assumption of asymmetrical causal ordering and allow a simple least-squares solution? Let us examine the following simultaneous equations in which we have represented each of $Z_1, Z_2, \ldots, Z_n$ (in standard-units) as a dependent variable with residuals $Z_a, Z_b, \ldots, Z_k$:

$$\begin{aligned}
Z_1 &= P_{12}Z_2 + P_{13}Z_3 + \cdots + P_{1n}Z_n + P_{1a}Z_a \\
Z_2 &= P_{21}Z_1 + P_{23}Z_3 + \cdots + P_{2n}Z_n + P_{2b}Z_b \\
&\vdots \\
Z_n &= P_{n1}Z_1 + P_{n2}Z_2 + \cdots + P_{n,n-1}Z_{n-1} + P_{nk}Z_k
\end{aligned} \tag{2.1}$$

This type of simultaneous equation structure will not suffice for our purposes, because it does not meet the two conditions stated above. First, it does not represent an asymmetrical causal system since it includes both the path coefficients P_{ij} and P_{ji}, for each i and j. Second, unless some of the path coefficients are set equal to zero, there is no set of values which yields a unique solution for the path coefficients, and least-squares procedures cannot be utilized to solve the system (Blalock, pp. 53-54).

We can solve both of the above difficulties by adjusting the system so as not to permit two-way causation. This implies that, if we allow for the possibility that $P_{ij} \neq 0$, the P_{ji} must be zero. In equations (2.1), let us set each P_{ij} equal to zero if $j > i$. This

condition gives the following type of simultaneous structure which is called a recursive system of simultaneous equations:

$$
\begin{aligned}
Z_1 &= P_{1a}Z_a \\
Z_2 &= P_{21}Z_1 + P_{2b}Z_b \\
Z_3 &= P_{31}Z_1 + P_{32}Z_2 + P_{3c}Z_c \\
&\cdot \\
&\cdot \\
&\cdot \\
Z_n &= P_{n1}Z_1 + P_{n2}Z_2 + \cdots + P_{n,n-1}Z_{n-1} + P_{nk}Z_k
\end{aligned}
\qquad (2.2)
$$

There are several important properties of recursive equations as illustrated by (2.2). First, note that Z_1 is taken to be caused only by variables that are outside of the set under consideration. Hence, Z_1 corresponds to what we have called an exogenous variable. However, Z_2 is causally dependent on Z_1 as well as a residual variable. Furthermore, Z_3 is causally dependent on both Z_1 and Z_2 and a residual variable. Finally, Z_n is dependent on all of the other Z_i and a residual variable. A second important property of recursive systems is that any of the remaining path coefficients may be set equal to zero if it does not reflect a postulated causal dependency. For example, if we set $P_{21} = 0$, then we have a representation of a multivariate system in which there are two exogenous variables - Z_1 and Z_2. Finally, perhaps the most important property of recursive systems of regression equations is that we can make use of ordinary least-squares procedures to estimate the postulated path coefficients (Wold and Jureen, pp. 51-52).

Let us now try to grasp some intuitive understanding of what the term "recursive" means when applied to a system of equations and build a rationale for an assumption about recursive equations which we will find convenient to make. First, note that, if we enter a recursive set of equations to determine the value of, say, Z_j, we will find it is dependent on the values of all Z_i for $i < j$. If we proceed to inquire into the determination of the values of each of these Z_i in turn, we will find that they are dependent on the variables which preceded them in the system until at last we come to the variable(s) the value of which is simply taken as given or observed and caused by no explicitly considered variable. Thus, the statement that a system of equations is recursive means that there is at least one variable the value of which is not in question and which successively enters into the determination of every other variable in the system either directly or indirectly through the determination of an intermediate variable. Now consider what happens in equation (2.2) if an increase in the value of Z_a is associated with an increase in the value of Z_b, i.e., if there is a positive correlation between "all other" variables which cause variation in Z_2 and those which cause variation in Z_1. Then, as Z_1 ($= Z_a$) increases, the value of Z_b will also increase. This will cause the estimate for P_{21} to be spuriously high in order to compensate for the confounding effect of Z_b. Similar types of side-effects can be traced through equations (2.2) if any of the residual terms are correlated. Hence, in order to assure that we get unbiased estimates of the path coefficients in path models involving recursive systems of equations, we shall have to assume that

the residual terms in each of the equations are uncorrelated. Practically, this means that, in a specific empirical problem, we shall want to bring as many as possible of the common components of the residuals explicitly into the path model so that the residuals will have or approximate zero correlation. Furthermore, we shall find that we can test this assumption in certain models.

Section 3: Path Analysis Revisited.

3.1. The Multi-stage, Multivariate Path Model. Consider again the problem posed at the beginning of Section 2. Briefly, the type of system we would like to handle concerns "stages" or "chains" of causation. The notion of recursive systems of equations provides a tool for the development of a general type of model for such multivariate systems which we shall term the multi-stage, multivariate path model. Because there are so many possible specific uses of this model, it is virtually impossible to discuss it in general. Therefore, we shall utilize an example from a sociological problem posed to the author by David C. Eaton. Eaton was concerned with the explanation of the personal income of heads of households in the United States in the year 1959 by a limited number of personal characteristics of the heads. From a search of relevant literature, he was able to find a number of bivariate correlations among his variables of interest. Furthermore, on the basis of time sequences and theoretical assumptions, he was willing to postulate a causal ordering among the variables. Specifically, he was interested in explaining the personal income of heads of households from their personal character-

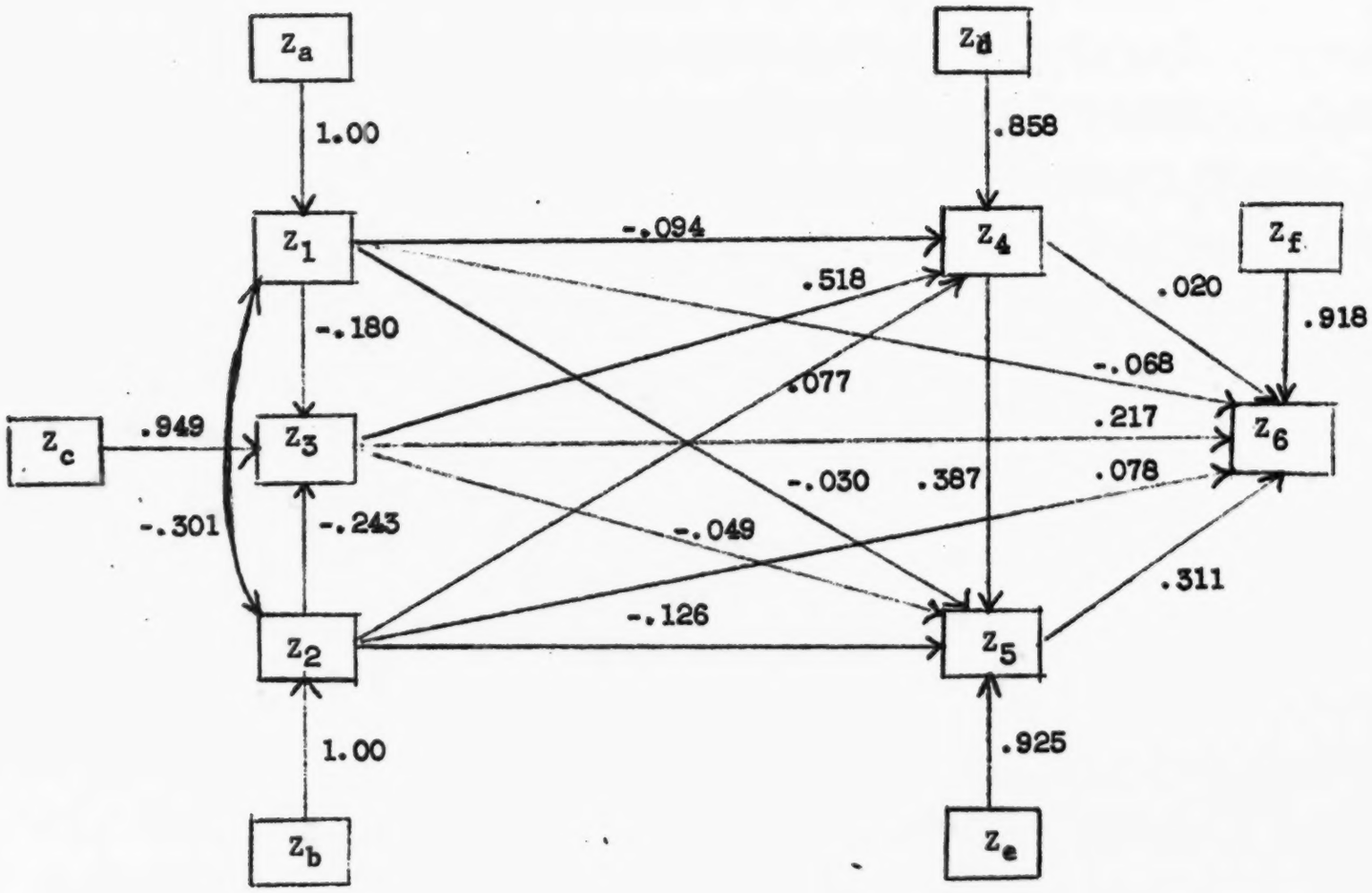
istics of (1) race, (2) age, (3) education, (4) occupation, and (5) fullness-regularity of employment. Since he was not willing to consider each of these variables as exogenous with no causal relations to or from the others, the elementary multivariate path model was obviously not the appropriate model. For example, because of its priority in time (being determined at birth), race was taken as an exogenous variable which generally has had an asymmetrical causal effect on the level of education, status of occupation, and employment-fullness-regularity of heads of households through institutionalized patterns of racial discrimination throughout the society. On the other hand, race was postulated to have only a symmetric non-causal relationship to age which, in turn, was taken as an exogenous variable having an asymmetrical effect, first of all, on education because of the differing levels of educational experience of each age cohort (generation). Age was further postulated to have an asymmetric positive effect on occupational status of heads of household through institutionalized patterns of seniority, tenure, and promotion. This relationship was also expected because of the exclusion of heads of households 65 or more years old from the sample - an age at which the relationship would be expected to become negative for a number of reasons. Finally, age was expected to have direct effects on employment-fullness-regularity and personal income of heads of households. Education of heads of households was taken as the first of the dependent variables. However, in addition to its postulated dependency on race and age, education was itself taken to have direct effects on the occupation, employment-fullness-regularity and personal income of the heads

from a consideration of institutionalized patterns of hiring and employment in industrial societies. Again, for institutional reasons, the status of a head's occupation was taken to be dependent on his race, age, and education while, in turn, occupation was postulated to have an asymmetrical effect on the fullness and regularity of employment and income of heads. Finally, employment-fullness-regularity was posited as dependent on the other variables and as determining income. A path diagram for this complex pattern of relationships is given in Figure 6 with the numerical values of the postulated path and correlation coefficients.

Having mapped the 6 variables onto a path diagram representing the rough notions of causation with which we began, it was quite simple to write the following recursive system of regression equations as the path model:

$$
\begin{aligned}
Z_1 &= P_{1a}Z_a \\
Z_2 &= P_{1b}Z_b \\
Z_3 &= P_{32}Z_2 + P_{31}Z_1 + P_{3c}Z_c \\
Z_4 &= P_{43}Z_3 + P_{42}Z_2 + P_{41}Z_1 + P_{4d}Z_d \\
Z_5 &= P_{54}Z_4 + P_{53}Z_3 + P_{52}Z_2 + P_{52}Z_1 + P_{5e}Z_e \\
Z_6 &= P_{65}Z_5 + P_{64}Z_4 + P_{63}Z_3 + P_{62}Z_2 + P_{61}Z_1 + P_{6f}Z_f
\end{aligned}
\qquad (3.1)
$$

Furthermore, with the aid of both the path diagram and the path model, twenty-three specific predictions regarding direct and indirect effects of the variables were deduced. All but two null hypotheses regarding these propositions were rejected. The reader is referred to Eaton (pp. 105-117) for complete details and evaluation.



where:

z_1 = Race of Head

z_2 = Age of Head

z_3 = Education of Head

z_4 = Occupation of Head

z_5 = Employment-Fullness-Regularity of Head

z_6 = Personal Income of Head

FIGURE 6.

The results for multivariate path models regarding computation of residual path coefficients and indirect effects hold also for this type of model. However, one must calculate a residual path coefficient for each equation in the path model by using the multiple correlation coefficient for that equation. As before, the correlation between any two variables of the model may be expanded along the lines of formula (1.20). Let us explore the correlation of Z_3 and Z_5 as an example:

$$\begin{aligned} r_{53} &= (1/N)\ \Sigma\ Z_3 Z_5/N \\ &= (1/N)\ \Sigma\ Z_3\ (P_{54}Z_4 + P_{53}Z_3 + P_{52}Z_2 + P_{51}Z_1 + P_{5d}Z_d) \\ &= P_{54}r_{34} + P_{53} + P_{52}r_{23} + P_{51}r_{13} \end{aligned} \tag{3.2}$$

The general form of this expansion theorem for multi-stage, multivariate path models is

$$r_{ij} = \sum_k P_{ik}r_{jk} \tag{3.3}$$

where i and j denote two variables in the system and the index k runs over all variables from which paths lead directly to Z_i. If we continue to expand (3.2) by means of (3.3), we have

$$\begin{aligned} r_{53} &= P_{53} + P_{54}r_{34} + P_{52}r_{23} + P_{51}\ (1/N\ \Sigma\ Z_1 Z_3) \\ &= P_{53} + P_{54}r_{34} + P_{52}r_{23} + P_{51}\ 1/N\ \Sigma\ Z_1\ (P_{32}Z_2 + \\ &\qquad P_{31}Z_1 + P_{3c}Z_c) \\ &= P_{53} + P_{54}r_{34} + P_{52}r_{23} + P_{51}P_{32}r_{12} + P_{51}P_{31} \\ &= P_{53} + P_{54}r_{34} + P_{52}\ (1/N\ \Sigma\ Z_2 Z_3) + P_{51}P_{32}r_{12} + P_{51}P_{31} \\ &= P_{53} + P_{54}r_{34} + P_{52}\ 1/N\ \Sigma\ Z_2\ (P_{32}Z_2 + P_{31}Z_1 + P_{3c}Z_c) \\ &\qquad + P_{51}P_{32}r_{12} + P_{51}P_{31} \end{aligned}$$

$$= P_{53} + P_{54}r_{34} + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}(1/N \Sigma Z_3Z_4) + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}\,1/N \Sigma Z_3 (P_{43}Z_3 + P_{42}Z_2 + P_{41}Z_1 + P_{4d}Z_d) + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}P_{43} + P_{54}P_{42}r_{23} + P_{54}P_{41}r_{13} + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}P_{43} + P_{54}P_{42}(1/N \Sigma Z_2Z_3) + P_{54}P_{41}(1/N \Sigma Z_1Z_3) + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}P_{43} + P_{54}P_{42}\,1/N \Sigma Z_2 (P_{32}Z_2 + P_{31}Z_1 + P_{3c}Z_c) + P_{54}P_{41}\,1/N \Sigma Z_1 (P_{32}Z_2 + P_{31}Z_1 + P_{3c}Z_c) + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31}$$

$$= P_{53} + P_{54}P_{43} + P_{54}P_{42}P_{32} + P_{54}P_{42}P_{31}r_{12} + P_{54}P_{41}P_{32}r_{12} + P_{54}P_{41}P_{31} + P_{52}P_{32} + P_{52}P_{31}P_{21} + P_{51}P_{32}r_{12} + P_{51}P_{31} \qquad (3.4)$$

This type of expansion can be carried out for all of the correlations between any two variables in the model. It may yield valuable information

regarding indirect effects. If we subtract P_{53} from both sides of (3.4), then we will have the familiar formula for the computation of the total indirect effect of variable 3 on variable 5. Furthermore, in some empirical problems we may want to examine <u>each of the separate indirect effects</u> as on the right-hand side of (3.4). An unusual finding regarding the relationship of status of occupation (Z_4) to personal income of of head (Z_6) in the present example was that its direct effect was extremely small (0.020) when employment-fullness-regularity was controlled. However, the indirect effect of occupation on income through its <u>direct</u> effect on employment-fullness-regularity (Z_5) and the <u>direct</u> effect of (Z_5) on income, i.e., the product (P_{54}) (P_{65}), was +0.238. We must emphasize to the point of becoming polemical that it is through such detailed findings as this that we will move behavioral research and theory beyond the quagmire of simplistic bivariate propositions to the realm of multivariate knowledge!

At this point, we wish to reiterate that the model of Figure 6 is only one example of a possible infinity of specific multi-stage, multivariate path models. However, it suffices to illustrate the general principles to follow when dealing with a complex pattern of dependent relationships.. From his experience with such problems, the author suggests <u>the following steps as a general procedure.</u> <u>First</u>, since the researcher often approaches multi-stage, multivariate problems with only crude ideas of the proper causal structure to postulate, he should begin to formalize his notions by mapping them onto a path diagram which he may use as a heuristic device until he is satisfied that it represents

the causal sequences as suggested by the current state of theoretical and empirical knowledge about the variables of interest. In general, the multi-stage, multivariate path model may include any number of exogenous variables and any number of causal stages with any number of dependent variables at each stage. Furthermore, as emphasized in the discussion of recursive systems, path coefficients from all preceding variables in the model need not be postulated for subsequent variables - note that in Figure 6 all path coefficients were postulated - if there is some theoretical or empirical reason for postulating that they will be zero or near-zero. Of course, if one or more path coefficients is predicted to be zero, then the researcher should run the model both with and without those path coefficients to ascertain whether or not they actually disappear in the empirical data. Second, if the researcher is satisfied with the structure represented by his path diagram, then he should write the path model or set of recursive equations which is implied by the diagram. This set of equations constitutes the actual regressions from which he gets estimates of the postulated path and correlation coefficients. Third, he should compute the residual path coefficients by applying formula (1.23) to each of the equations in the path model. Fourth, the researcher should compute estimates of total indirect effects of prior variables on subsequent variables from formula (1.21). Fifth, if he is interested in probing in detail the manner in which a prior variable effects a subsequent variable, then he is encouraged to expand the correlation between the two variables along the lines of formula (3.3) and as illustrated by equation (3.4).

This five-fold procedure should facilitate the representation and interpretation of the most complex sequences of causal dependency.

3.2. The Multi-stage, Bivariate Path Model. If we postulate a series of causal stages but restrict the number of variables at each stage to two measured variables and a residual, then we have a special case of the multi-stage, multivariate path model which we may refer to as the multi-stage, bivariate path model or simple causal chain. It is instructive to examine this particular model because of the opportunity it provides to test the basic assumption of recursive systems of regression equations, viz., that the residual terms of the equations are uncorrelated.

For purposes of illustrating this model, we shall use the computations and data reported by Duncan (pp. 10-12). He postulated a multi-stage, bivariate path model to account for recently reported correlations between the occupational prestige ratings of four studies completed at widely separated dates: Counts (1925), Smith (1940), National Opinion Research Center (1947), and NORC replication (1963). The path model postulated by Duncan is given by the recursive system:

$$\begin{aligned} Z_1 &= P_{1a}Z_a \\ Z_2 &= P_{21}Z_1 + P_{2b}Z_b \\ Z_3 &= P_{32}Z_2 + P_{3c}Z_c \\ Z_4 &= P_{43}Z_3 + P_{4d}Z_d \end{aligned} \qquad (3.5)$$

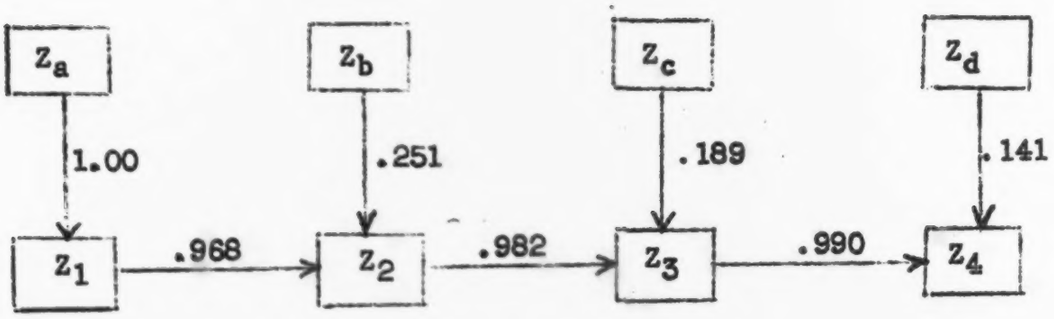
where Z_1 = Counts prestige ratings, 1925, Z_2 = Smith prestige ratings, 1940, Z_3 = NORC prestige ratings, 1947, Z_4 = NORC prestige ratings, 1963, and Z_a, Z_b, Z_c, Z_d are residual variables. A path diagram of the postu-

lated model with numerical values for the path coefficients is given in Figure 7(a).

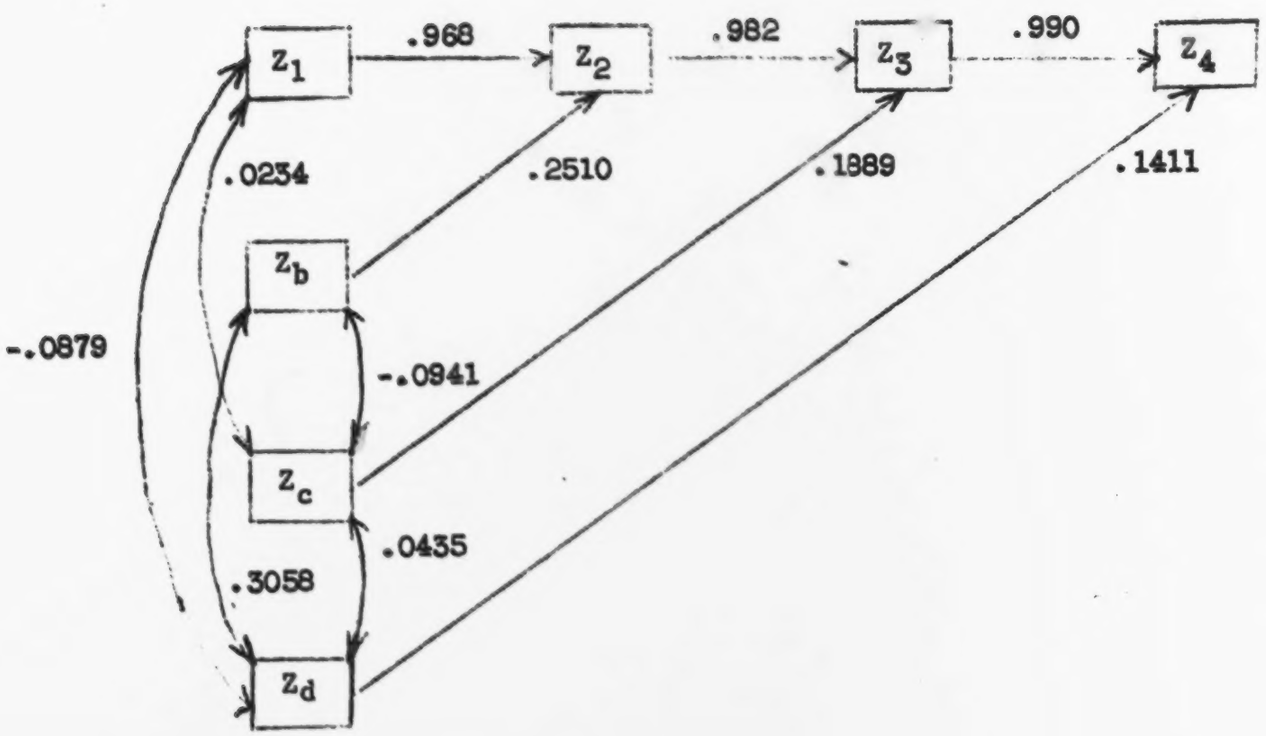
Table 3 gives the estimators of the path coefficients for the simple causal chain of Figure 7(a). As was shown in section 1.3 and illustrated in formulas (1), (2), and (3) of Table 3, the estimators for path coefficients in the bivariate case are correlation coefficients. Furthermore, equations (4), (5), and (6) of Table 3 show that the computation of the residual path coefficient is an immediate result of the formula for complete determination of the dependent variable. However, up until now we have not questioned the assumption of recursive systems that the residuals are uncorrelated. Formulas (7), (8), and (9) illustrate a condition which must be met by simple causal chains if that assumption is tenable. That is, if the assumption of uncorrelated residuals holds, then the <u>observed</u> correlation coefficient of the alternate or terminal variables of Figure 7(a) must equal the product of the <u>observed</u> path coefficients connecting them. Duncan (p. 11) gives the following calculations for Figure 7(a):

Variables	Calculated Correlation	Observed Correlation	Difference
r_{31}	.951	.955	.004
r_{42}	.972	.971	-.001
r_{41}	.942	.934	-.008

Although the discrepancies between inferred and observed correlations are small and trivial enough so that we may accept the hypothesis of a multi-stage, bivariate path model with uncorrelated residuals, Duncan,


z_a
z_b
z_c
z_d
1.00
.251
.189
.141
z_1
.968
z_2
.982
z_3
.990
z_4

(a)


z_1
.968
z_2
.982
z_3
.990
z_4
.0234
.2510
.1889
.1411
z_b
-.0879
-.0941
z_c
.0435
.3058
z_d

(b)

FIGURE 7.

TABLE 3: ESTIMATORS OF PATH COEFFICIENTS FOR THE PATH MODEL OF FIGURE 7(a).

(1)	r_{21}	=	P_{21}
(2)	r_{32}	=	P_{32}
(3)	r_{43}	=	P_{43}
(4)	r_{22}	= 1 =	$P_{21}^2 + P_{2b}^2$
(5)	r_{33}	= 1 =	$P_{32}^2 + P_{3c}^2$
(6)	r_{44}	= 1 =	$P_{43}^2 + P_{4d}^2$
(7)	r_{31}	=	$P_{21}P_{32}$
(8)	r_{42}	=	$P_{32}P_{43}$
(9)	r_{41}	=	$P_{21}P_{32}P_{43}$

to illustrate what should be done in case the discrepancies had been large, constructed the alternative model shown in Figure 7(b). In this model, Duncan has dropped the assumption of uncorrelated residuals and computed the correlations among them which must be postulated to explain the small discrepancies between the inferred and observed path coefficients discussed above. However, the assumption that a residual is uncorrelated with the immediately preceding variable in the chain holds for 7(b). The formulas provided by Duncan (p. 12) which yield the desired coefficients when solved in order are given in Table 4.

In general, as Duncan points out, if we are considering a k-variable causal chain, we must estimate k-1 residual paths (3 for Figure 7(b)), (k-1) (k-2)/2 correlations between residuals (3 for Figure 7(b)), k-1 paths for links in the chain (3 for Figure 7(b)), and k-2

TABLE 4: ESTIMATORS OF PATH COEFFICIENTS FOR THE PATH MODEL OF FIGURE 7(b).

(1)	r_{21}	=	.968	
(2)	r_{32}	=	.982	
(3)	r_{43}	=	.990	
(4)	r_{22}	= 1 =	P_{21}^2	+ P_{2b}^2
(5)	r_{33}	= 1 =	P_{32}^2	+ P_{3c}^2
(6)	r_{44}	= 1 =	P_{43}^2	+ P_{4d}^2
(7)	r_{31}	= .955 =	$P_{21}P_{32}$	+ $P_{3c}r_{c1}$
(8)	r_{41}	= .934 =	$P_{43}P_{32}P_{21}$	+ $P_{43}P_{3c}r_{c1}$ + $P_{4d}r_{1d}$
(9)	r_{42}	= .971 =	$P_{43}P_{32}$	+ $P_{21}P_{4d}r_{1d}$ + $P_{2b}P_{4d}r_{bd}$
(10)	r_{c2}	= 0 =	$P_{2b}r_{bd}$	+ $P_{21}r_{c1}$
(11)	r_{d3}	= 0 =	$P_{3c}r_{cd}$	+ $P_{32}r_{2d}$
	(where r_{2d}	=	$P_{21}r_{1d}$	+ $P_{2b}r_{bd}$)

correlations between the initial variable and residuals a,b,...,j in the chain (2 for Figure 7(b)). This yields a total of $(k^2 + 3k - 6)/2$ quantities to be estimated. For the purpose of estimation, we may construct $k(k-1)/2$ equations expressing known correlations in terms of paths (as in equations (1), (2), (3), (7), (8), and (9) of Table 4), k-1 equations of complete determination (as in equations (4), (5), and (6) of Table 4) and k-2 equations in which the correlation of a residual with the immediately preceding variable in the chain is set equal to zero (equations (10) and (11) of Table 4). This gives $(k^2 + 3k - 6)/2$ equations, the precise number needed for a unique solution.

The procedure for testing the assumption of uncorrelated residuals, as illustrated by Duncan's example, may be useful in exploring relationships among variables which have been traditionally assumed to form simple causal chains. If, after the uncorrelated residual assumption is abandoned, the empirical data are still not sufficiently accounted for, then the assumption of a simple causal chain should be relaxed. Duncan's comments (p. 12) regarding his example are particularly appropriate here:

> ... The solution may, of course include meaningless results (e.g., $r > 1.0$), or results that strain one's credulity. In this event, the chain hypothesis had best be abandoned or the estimated paths modified.
>
> In the present illustration, the results are plausible enough. Both the Counts and the Smith studies differed from the two NORC studies and from each other in their techniques of rating and sampling. A further complication is that the studies used different lists of occupations, and the observed correlations are based on differing numbers of occupations. There is ample opportunity, therefore, for correlations of errors to turn up in a variety of patterns, even though the chain hypothesis may be basically sound. We should observe, too, that the residual factors here include not only extrinsic disturbances but also real though temporary fluctuations in prestige, if there be such.
>
> What should one say, substantively, on the basis of such an analysis of the prestige ratings? Certainly, the temporal ordering of the variables is unambiguous. But whether one wants to assert that an aspect of social structure (prestige hierarchy) at one date 'causes' its counterpart at a later date is perhaps questionable. The data suggest there is a high order of persistence over time, coupled with a detectable, if rather glacial, drift in the structure. The calculation of numerical values for the model hardly resolves the question of ultimate 'reasons' for either the pattern of persistence or the tempo of change. These are, instead, questions raised by the model in a clear way for further discussion and, perhaps, investigation.

3.3. The Path Decomposition Model. Because many of the dependent variables of interest in the behavioral sciences are composite variables, we now discuss a use of path analysis which may be referred to as the path decomposition model. Thus, various tests or scales commonly used in behavioral research are composed of subscales or subtests. In the case of such composite dependent variables, it is often of interest (1) to compute the relative contributions of the component variables to variation in the composite variable, and (2) to ascertain how independent variables effect the composite variable via its components.

We shall again utilize an illustrative example provided by Duncan (pp. 7-10). However, since the subject matter of his example is rather specific (population density), we shall discuss his model symbolically; the form of his composite variable has wide generality. The raw-score definition of the composite variable V_0 is

$$V_0 = V_1 \cdot V_2 \cdot V_3$$

Let $X_0 = \log V_0$, $X_1 = \log V_1$, $X_2 = \log V_2$, and $X_3 = \log V_3$. Then the composite variable is an additive combination of X_1, X_2, and X_3:

$$X_0 = X_1 + X_2 + X_3$$

If each variable is expressed in standard-score form, we may write

$$Z_0 = P_{01}Z_1 + P_{02}Z_2 + P_{03}Z_3 \qquad (3.6)$$

as the path decomposition model where $Z_0, \ldots, Z_3$ are the variables in standard-score form and P_{01}, P_{02}, P_{03} are the path coefficients involved

in the determination of Z_0 by Z_1, Z_2, and Z_3. In the case of complete determination by measured variables, the definition of the path coefficient, as given by formula (1.4), reduces to a ratio of standard deviations since the partial regression coefficient part of the definition (c_{ij}) is unity. Hence, the following numerical values of the path coefficients, as Duncan indicates (p.8), may be computed without prior calculation of correlations:

$$P_{01} = S_1/S_0 = .132 \qquad S_0 = .491 \qquad S_1 = .065$$
$$P_{02} = S_2/S_0 = .468 \qquad\qquad\quad S_2 = .230$$
$$P_{03} = S_3/S_0 = .821 \qquad\qquad\quad S_3 = .403$$

The path diagram provided by Duncan (p. 9) for the present path model is given in Figure 8(a). Table 5 gives Duncan's correlation matrix for the present problem. The composition of the correlation of the composite variable with its component parts may now be written from formula (3.3).

$$r_{01} = P_{01} + P_{02}r_{12} + P_{03}r_{13} = -.419$$
$$r_{02} = P_{01}r_{12} + P_{02} + P_{03}r_{23} = .636$$
$$r_{03} = P_{01}r_{13} + P_{02}r_{23} + P_{03} = .923$$

As Duncan points out (p. 8), this preliminary analysis gives a clear ordering of the three components in terms of relative importance, as indicated by the path coefficients, and shows that one of the components is actually negatively correlated with the composite variable, because of its negative correlations with the other two components.

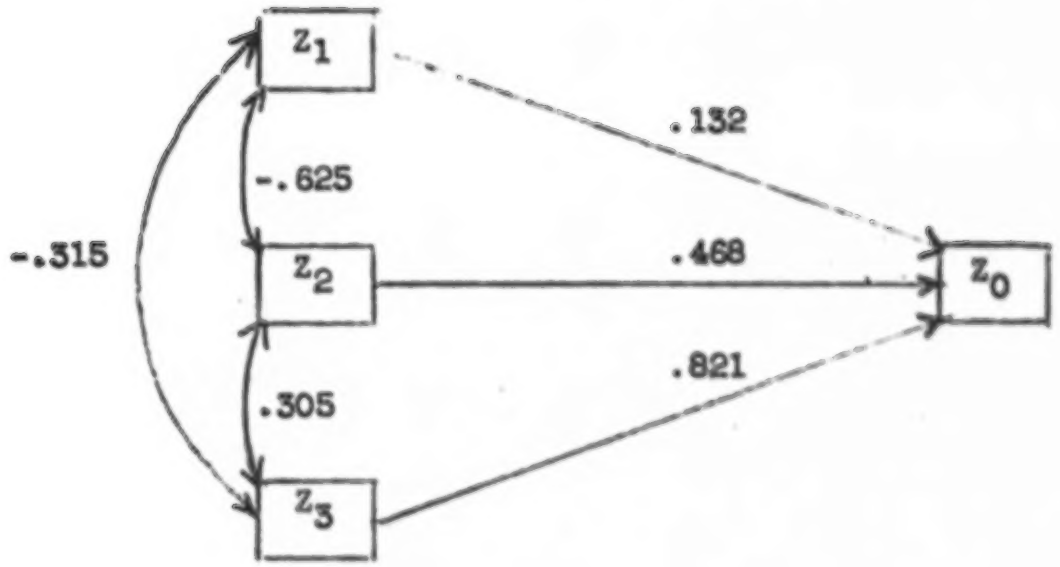
TABLE 5: CORRELATION MATRIX FOR LOGARITHMS OF VARIABLES IN DUNCAN'S PATH DECOMPOSITION EXAMPLE

Variable	Z_1	Z_2	Z_3	Z_4	Z_5
Z_0	-.419	.636	.923	-.663	-.390
Z_1		-.625	-.315	.296	.099
Z_2			.305	-.594	-.466
Z_3				-.517	-.226
Z_4					.549

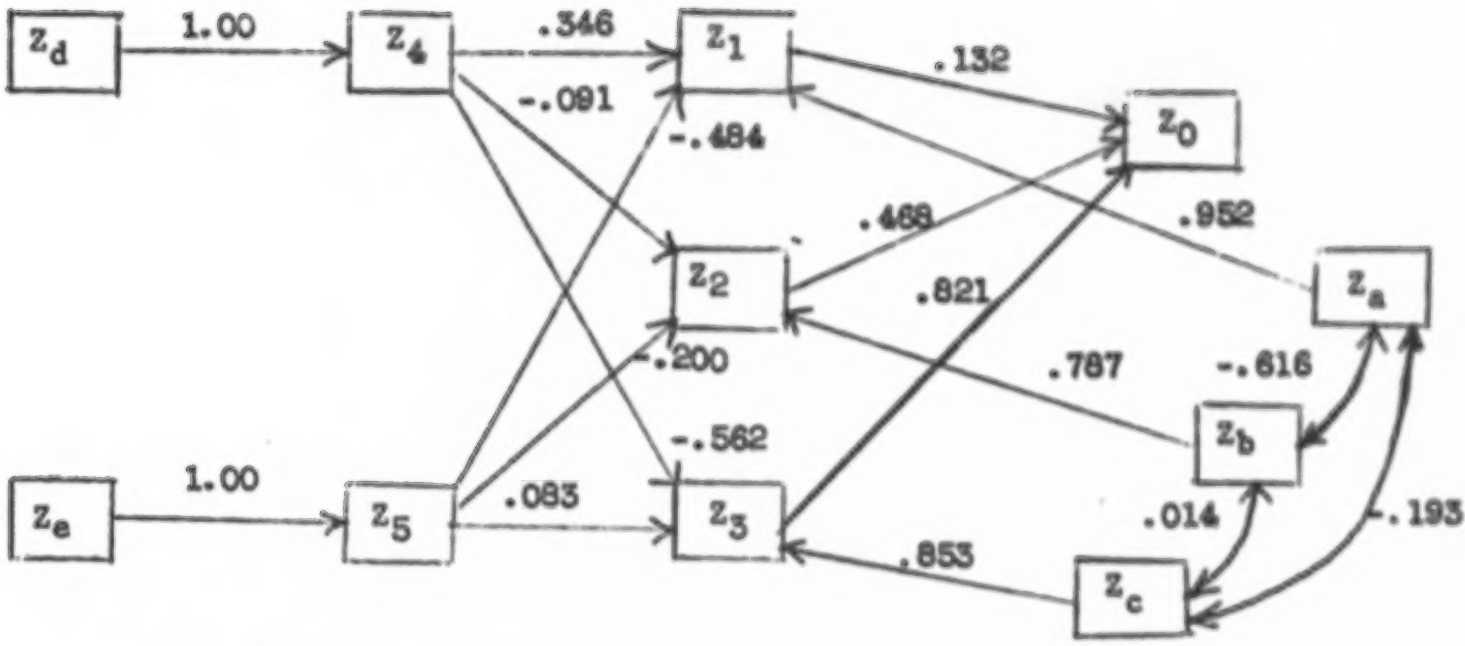
Duncan postulates a second path model to account for the relationship of the composite variable to two independent variables via its components. A path diagram for this model is given in Figure 8(b). The path model is:

$$
\begin{aligned}
Z_4 &= P_{4d}Z_d \\
Z_5 &= P_{5e}Z_e \\
Z_3 &= P_{34}Z_4 + P_{35}Z_5 + P_{3c}Z_c \\
Z_2 &= P_{24}Z_4 + P_{25}Z_5 + P_{2b}Z_b \\
Z_1 &= P_{14}Z_4 + P_{15}Z_5 + P_{1a}Z_a \\
Z_0 &= P_{01}Z_1 + P_{02}Z_2 + P_{03}Z_3
\end{aligned}
\qquad (3.7)
$$

It should be noted that zero correlation is not assumed in this model for the residuals Z_a, Z_b, and Z_c. The path coefficients - P_{34}, P_{35}, P_{24}, P_{25}, P_{14}, and P_{15} - are standardized partial regression coefficients as for other multivariate path models. Furthermore, the residual path coefficients are given by formula (1.23). Duncan's discussion of the


z_1
z_2
z_3
z_0
.132
-.625
-.315
.468
.821
.305

(a)


z_d
1.00
z_4
.346
z_1
.132
-.091
-.484
z_0
.952
.468
z_2
z_a
.821
-.200
.787
-.616
-.562
z_b
z_e
1.00
z_5
.083
z_3
.014
.193
.853
z_c

(b)

FIGURE 8.

computation of the residuals (pp. 9-10) is appropriate (using the symbols of this paper):

> The two independent variables by no means account for all the variation in any of the components, as may be seen from the size of the residuals, P_{1a}, P_{2b}, and P_{3c},... It is possible, nevertheless, for the independent variables to account for the intercorrelations of the components, and, ideally, one would like to discover independent variables which would do just that. The relevant calculations concern the correlations between residuals. These are obtained from the basic theorem, equation [3.3], by writing, for example,
>
> $$r_{23} = P_{24}r_{34} + P_{25}r_{35} + P_{2b}P_{3c}r_{bc},$$
>
> which may be solved for $r_{bc} = .014$. In this setup, the correlations between residuals are merely the conventional second-order partial correlations; thus $r_{ab} = r_{12.45}$, $r_{ac} = r_{13.45}$, and $r_{bc} = r_{23.45}$. Partial correlations, which otherwise have little utility in path analysis, turn out to be appropriate when the question at issue is whether a set of independent variables 'explains' the correlation between two dependent variables. In the present example, while $r_{23} = .305$, we find $r_{bc} = r_{23.45} = .014$. Thus the correlation between ... Z_2 and... Z_3 is satisfactorily explained by the respective relationships of these two components to ... Z_4 and ... Z_5. The same is not true of the correlations involving ... Z_1, but fortunately this is by far the least important component....

Finally, Duncan gives the following equations as the most compact answer to the question of how the effects of the independent variables are transmitted to the dependent variable via its components (p.10):

$$r_{04} = P_{01}r_{14} + P_{02}r_{24} + P_{03}r_{34} \qquad (3.8)$$
$$= .039 - .278 - .424 = -.663$$

and

$$r_{05} = P_{01}r_{15} + P_{02}r_{25} + P_{03}r_{35} \qquad (3.9)$$
$$= .013 - .218 - .185 = -.391$$

From these results, we note that the composite variable is negatively related to both independent variables, but the effects via the first component, although small, are positive. Furthermore, the relative importance of the effects of Z_4 and Z_5 via the second and third components is reversed in the two equations. Finally, a more detailed examination of the transmission of effects can be obtained by expansion of (3.8) and (3.9) by formula (3.3).

3.4. The Basic Assumptions of Path Analysis and Model Testing. At this point, it is appropriate to discuss the basic assumptions of path analysis, answer possible objections to the method, and mention extensions of the procedure which we have not developed in this paper.

The Assumptions of Linearity and Additivity. The assumption of linear, additive relationships among the variables of interest is made in most applications of correlation and regression analysis. Past empirical research in a particular problem area is a good basis for judging the tenability of the linearity and additivity assumptions. Otherwise, there are simple procedures such as point-plotting to explore the degree to which the assumptions hold in a specific set of data. Furthermore, although a relationship may not be linear throughout its entire range of values, it may be linear within the range of values under consideration. Finally, there are a number of convenient transformations, such as the logarithmic transformation of section 3.3, which may be used to transform data in order to meet the linearity and additivity assumptions. The important point to emphasize is that these are assumptions which are specific to the particular area of investigation and must be

not ordinarily willing to assume that a change in a variable caused a prior change in another variable. A second source of information for the causal ordering of the variables may be existing experimental or case-study results. Finally, the theoretical assumptions of the particular substantive area provide a third source forthe asymmetry assumption. Following Wright (quotation in section 1), we view a causal assumption as less of an assertion about empirical reality than as a strategy for inquiry. Path analysis, by itself, cannot prove the validity of a set of causal assumptions. It can only give the consequences of an assumed causal sequence for a set of data. It is, of course, the responsibility of the researcher to defend his causal assumptions in a given empirical study.

A particular type of behavioral science relationship for which the asymmetric path model may not be appropriate is the so-called "interdependent" relationship. That is, we often encounter variables such that a change in one causes a change in another which, in turn, leads to a change in the former, etc. The relationship of level of education and racial discrimination provides a common sociological example. At an original time of observation, some element in the social system may cause an increase in the level of Negro education. This increase in level of education may cause a decrease in racial discrimination. Furthermore, the decrease in racial discrimination may feedback through a number of mechanisms to cause another increase in the opportunity for, and level of,Negro education. Thus, the cycle may continue until an equilibrium position of the variables is established. How can we handle this type of relation-

ship with asymmetric path models? The answer would seem to depend on the rapidity of the feedback relationship. In the case of a relatively slow feedback process, our asymmetric path models will give us a crude, but perhaps meaningful, cross-section snapshot of the system which only approximates the "real" causal structure. Furthermore, the path model will have to be continuously updated over time as the regression relationships change. For those systems of relationships in which the feedback is relatively rapid, however, our asymmetric path models must be modified. Wright (1960b) has indicated how this may be done in certain specific types of problems. Economists also, with the problem of the rapid adjustment of market forces, have dealt at considerable length with this type of model. The present author intends to add a special section on feedback models to this paper in the near future.

The Testing of Models. Because of our emphasis on the representational and interpretive uses of path models, we have not addressed ourselves to the problems of tests of significance for path coefficients and testing procedures for alternative models in this paper. These topics also demand treatment in a special section. As a rule of thumb, however, we propose, particularly for small samples, that a criterion of at least 0.10 be set for the retention of a path coefficient in a model. This means that the variable must account for at least 10 per cent of the standard deviation, and 1 per cent of the variance, of the endogenous variable. A criterion of this level should not lead to the premature rejection of too many important variables.

ADDITIONAL READING

The topics discussed in this paper are referenced in a wide variety of publications in biometrics, econometrics, and statistics. If the reader desires to read their treatment in other sources, he may begin with the publications listed in the bibliography. Any of the articles by Wright are worthwhile discussions of path analysis. However, his June 1960 Biometrics article is a particularly good summary treatment. The articles by Duncan, Kempthorne, Tukey, and Turner and Stevens also provide good treatments of path analysis from somewhat different perspectives.

Recursive systems of equations are discussed on an elementary level in the book by Blalock. However, a somewhat more technical treatment is given in the Simon and Wold and Jureen references.

APPENDIX I

THE MATHEMATICS OF BIVARIATE CORRELATION AND REGRESSION

In what follows, we have not sought to develop a distinct population theory and a distinct sample theory of correlation and regression. We deal primarily with sample theory and utilize population theory in a heuristic fashion. Consider the case in which we have observations on two continuous, intervally-measured variables X and Y in raw-score form. Then we may define the correlation coefficient as follows:

Definition A.1.1. Suppose that X and Y are continuous, intervally-measured variables. Then the statement that r is the correlation coefficient of the observed values of the X and Y variables means that r is a number and that r is a measure of the degree of linear covariation of the observed values of X and Y such that

$$r = \frac{\Sigma (X-M_x)(Y-M_y)}{N\ S_x\ S_y} = \frac{\Sigma\ x\ y}{N\ S_x\ S_y} \qquad \text{(A1.1)}$$

where X and Y are the observed values of the variables in raw-score form, M_x and M_y are the means of the raw scores for the X and Y variables, S_x and S_y are the standard deviations of the observed values of X and Y, and x and y are the observed values of the X and Y variables in deviation-score form, i.e., $x = X-M_x$ and $y = Y-M_y$.

Now we would like to develop a means for predicting Y values for given X values and for interpreting the correlation coefficient. We begin by assuming that the X and Y variables are linearly related, i.e., we assume that the functional form of the relationship of Y and X in the

population of X and Y values is a line of the form $Y = A'' + B''X$, where A" is the Y intercept (value of Y where the line crosses the Y axis) and B" is the slope of the line (the inclination of the line to the X axis).

At this point, then, our problem is merely to develop estimators of the A" and B" coefficients from the observed X and Y values. Of the several methods of estimation, we shall follow traditional practice here and adopt as a basis for an estimated best-fit line the criterion that the sum of the squares of the deviations from the line shall be as small as possible. In symbols, let $Y' = A + BX$, where Y' (read Y-prime) is the value estimated, for a given X, of the Y variable, A is the estimated value for A", B is the estimated value for B", and let Y be the observed value of the Y variable. Then $(Y-Y')^2$ represents the squared deviation of any Y from the estimated value. Our problem is to choose the estimators A and B so as to make $\Sigma(Y-Y')^2$ as small as possible. We shall find it more convenient to deal with both the equation, $y' = a + bx$, and the sum, $\Sigma(y-y')^2$, in deviation-units, with y' and y as deviations from M_y and $x = X-M_x$. This is merely a translation of the reference axes to make the origin coincide with M_x and M_y. Therefore, the value of a in the equation becomes zero and we shall drop it from further consideration.

This allows us to write $y' = bx$ as the equation for estimating y, in deviation-units, from x, or deviation-values of x. Our problem now is that of determining the value of b which will make $\Sigma(y-y')^2$ a minimum. It shall be simple, once the optimal value for b has been determined, to pass back to the original reference frame, the gross-

score axes, by substituting for y' the values $Y'-M_y$, and for x, $X-M_x$.

The following derivation of the optimal value for b more or less follows that given by McNemar (pp. 122-3). We begin by setting up the function

$$f = \frac{\Sigma (y-y')^2}{N} = \frac{\Sigma (y-bx)^2}{N}$$

in which we have N deviations of the form y-y' or y-bx. The sum of these squared deviations divided by N gives us the function which we want to minimize by the proper choice of b. We shall choose the proper value of b by utilizing a theorem from the calculus. According to this theorem, we may minimizethe above function by taking its derivative with respect to b, setting this derivative equal to zero, and then solving for b. Thus

$$\frac{df}{db} = \frac{-2 \Sigma x(y-bx)}{N}$$

which, set equal to zero, divided by -2, yields

$$\frac{\Sigma x (y-bx)}{N} = 0$$

or

$$\frac{\Sigma xy - b\Sigma x^2}{N} = 0$$

then

$$\frac{\Sigma xy}{N} - b \frac{\Sigma x^2}{N} = 0$$

The first term involves the correlation coefficient as defined by formula (A1.1), from which definitional formula we see that $\Sigma xy/N = rS_xS_y$; and since $\Sigma x^2/N = S_x^2$, we have

$$rS_xS_y - bS_x^2 = 0$$

or

$$rS_y - bS_x = 0$$

which gives

$$b = r\frac{S_y}{S_x}$$

as the proper value for b. We therefore have

$$y' = r\frac{S_y}{S_x}x \qquad \text{(A1.2)}$$

as the equation for the best-fit line in deviation-score form. By proper substitution, we have

$$Y'-M_y = r\frac{S_y}{S_x}(X-M_x)$$

or

$$Y' = r\frac{S_y}{S_x}X + (M_y - r\,S_y\,M_x) \qquad \text{(A1.3)}$$

as the equation in terms of the original or raw-scores. It is this form which we would use in predicting Y from X. Note that $B = b = r(S_y/S_x)$ is the slope of the line and that the constant A is the term in parentheses. Furthermore, note that we can get another form of equation (A1.2) by dividing both sides of the expression by S_y:

$$\frac{y'}{S_y} = r \frac{x}{S_x}$$

The observant reader will recognize the y'/S_y and x/S_x terms as the X and (predicted) Y variables in standard-score form, or

$$Z'_y = r\ Z_x \qquad \text{(A1.4)}$$

or

$$Z'_y = B^*\ Z_x \qquad \text{(A1.5)}$$

where $B^* = r$. B^* is usually referred to as <u>the standardized regression coefficient or beta weight</u>. It is usually symbolized by the Greek letter Beta, but, since we do not have a Greek letter for Beta on our keyboard, we shall denote it by B^*. Thus, we have three formulas for the line of best fit - (A1.2), (A1.3), and (A1.5) - corresponding to deviation-, observed-, and standard-score units, respectively.

We note two additional relations. First, we derive a formula for $B_{yx} = b_{yx}$ - the slope of the least-squares equation estimating the regression of Y on X - in terms of deivation scores. From (A1.3), we have

$$B_{yx} = r \frac{S_y}{S_x}$$

$$= \frac{\Sigma xy/N}{S_x\ S_y} \frac{S_y}{S_x}$$

$$= \frac{\Sigma xy/N}{S_x^2} \qquad \text{(A1.6)}$$

Second, just as we can estimate Y values from X values we can extend our equations to the estimation of X values from Y values, although we may never desire to do this in practice. Then we may write the equation for estimating X from Y as follows

$$X' = A_{xy} + B_{xy} Y$$

where $A_{xy} = M_x - r \frac{S_x}{S_y} M_y$ and $B_{xy} = \Sigma xy/N = r \frac{S_x}{S_y}$. Then we may write the product

$$B_{yx} B_{xy} = \frac{(\Sigma xy)^2/N^2}{S_x^2 S_y^2} = r^2 \qquad \text{(A1.7)}$$

At this point, it is profitable to review three properties or interpretations of the correlation coefficient.

A.1.1. Rate of Change. It is obvious from equation (A1.4) that the correlation coefficient may be interpreted as a rate of change - the amount of change in variable Y per unit of change in variable X - in standard-score form. It may also be shown that the correlation coefficient has bounds of +1 and -1. Hence, the largest possible change in Y, given a standard deviation of change in X, is plus or minus one standard deviation. It should also be noted from equations (A1.2) and (A1.3) that, if the standard deviations of the X and Y variables are equal, then the correlation coefficient may be interpreted as a rate of change for the variables in deviation- or observed-score units.

A.1.2. Accuracy of Prediction. The next property or interpretation of the correlation coefficient concerns the accuracy of pre-

diction by means of the regression equation. That is, we would like to be able to set up confidence bands or intervals about the regression line or predicted y-values in which we can have confidence that most of the actual observations fall with a specified degree of probability. In order to accomplish this goal, we need a measure of dispersion for the regression line comparable to, say, the standard error of estimate of the mean of a distribution. By introducing an assumption we may derive such a measure. We must assume that the standard deviations of the distribution of Y values for each value of X are equal in the population from which the sample is drawn. This assumption of homoscedasticity implies that, if we had a much larger sample size, the standard deviations of Y-values for each X value would be very nearly equal.

Note that $y-y'$ (or $Y-Y'$) represents the discrepancy between estimated and observed values and that $\Sigma(y-y')^2/N$ is the mean of the squared deviations, the root of which will be the standard deviation of the discrepancies between estimated and observed values. This particular standard deviation shall be called the standard error of estimate of the regression of Y on X and shall be denoted by $S_{y.x}$. We may derive an algebraic form for this expression as follows. By definition,

$$S^2_{y.x} = \frac{\Sigma (Y-Y')^2}{N} = \frac{\Sigma (y-y')^2}{N}$$

but

$$y' = r \frac{S_Y}{S_X} x$$

from (A1.2), so

$$s^2_{y.x} = \frac{1}{N} \Sigma \left(y - r \frac{S_y}{S_x} x\right)^2$$

$$= \frac{1}{N} \Sigma \left(y^2 - 2r \frac{S_y}{S_x} xy + r^2 \frac{S^2_y}{S^2_x} x^2\right)$$

$$= \frac{\Sigma y^2}{N} - 2r \frac{S_y}{S_x} \frac{(\Sigma xy)}{N} + r^2 \frac{S^2_y}{S^2_x} \frac{(\Sigma x^2)^2}{N}$$

$$= s^2_y - 2r \frac{S_y}{S_x} r S_x S_y + r^2 \frac{S^2_y}{S^2_x} S^2_x$$

$$= s^2_y - r^2 s^2_y$$

then

$$S_{y.x} = S_y \sqrt{1 - r^2} \qquad \text{(A1.8)}$$

Hence, we have a second procedure for interpreting the correlation coefficient: in terms of the accuracy of prediction or closeness of fit of the regression line to the data. If no correlation exists, we see that the error of estimate is S_y. We should also note the term in (A1.8) which involves r i $\sqrt{1 - r^2}$. The expression $\sqrt{1 - r^2}$ is called the <u>coefficient of alienation</u>. Observe that if r = 0, its value is 1 and the error of estimate is S_y.

<u>A.1.3. Variance and Correlation.</u> It may be shown (see McNemar, p. 129) that the variance of a sum (or difference) of two independent variables is equal to the sum of their separate variances. Furthermore, it may be shown (see McNemar, p. 130) that the predicted y' and the residual (y-y') are independent. Therefore, we have $y = y' + (y-y')$ and, since the two parts are independent,

$$s^2_y = s^2_{y'} + s^2_{y.x} \qquad \text{(A1.9)}$$

where $S^2_{y.x}$ is the variance of the residuals, (y-y'). Dividing both sides of this equation by S^2_y, we get

$$1 = \frac{S^2_{y'}}{S^2_y} + \frac{S^2_{y.x}}{S^2_y} \qquad \text{(A1.10)}$$

from which we see that, since the two ratios add to unity either one can be interpreted as a proportion. In short, the ratio of $S^2_{y'}$ to S^2_y is the proportion of the variance in Y which can be predicted from X, and the ratio of $S^2_{y.x}$ to S^2_y represents the proportion of the variance of Y which is left over or remains or cannot be predicted from X. This is the same variance which results if we square formula (A1.8):

$$S^2_{y.x} = S^2_y (1 - r^2)$$

or

$$\frac{S^2_{y.x}}{S^2_y} = 1 - r^2$$

Hence, we may substitute this value in (A1.10)

$$1 = \frac{S^2_{y'}}{S^2_y} + 1 - r^2$$

from which we have the ratio

$$\frac{S^2_{y'}}{S^2_y} = r^2 \qquad \text{(A1.11)}$$

In short, the square of the correlation coefficient gives the proportion of the total variance of Y which is attributable to variance in X. Also, the proportion of variance in Y which is due to variables other than X is $1 - r^2$. This is a third possible interpretation of r.

APPENDIX II

THE MATHEMATICS OF MULTIVARIATE CORRELATION AND REGRESSION

We shall now extend our discussion to the multivariate case in which we attempt to predict one variable by using several other variables and to analyze its variance into component parts. We shall find some similarities to the bivariate case, but we shall also encounter some significant differences. Again, parts of our derivations more or less follow those of McNemar (pp. 169-178).

A.2.1. The Three-Variable Case. Consider first the problem of predicting X_1 (criterion) from a knowledge of X_2 and X_3 (predixtors). Geometrically, we can imagine this as involving three reference axes instead of two as in the bivariate case. Here we can think of the vertical axis as representing X_1 and the two horizontal axes as representing X_2 and X_3. We begin by assuming that the relationship of X_1 to X_2 and X_3 is linear, i.e., we assume that the functional form of the relationship of the variables in the population of X_1, X_2, and X_3 values is of the form $X_1 = A'' + B''_2 X_2 + B''_3 X_3$ in which A'' is the (population) value of X_1 where the plane representing the predicted values of X_1 cuts the X_1 axis (where X_2 and X_3 are zero), and B''_2 and B''_3 are the inclinations (population values) of the plane of predicted X_1 values to the X_2 and X_3 axes, respectively (the expected changes in X_1, given a unit of change in X_2 and X_3, respectively).

As in the bivariate case, our problem is to estimate A'', B''_2, and B''_3. Again, this is a least-squares affair - the sum of the squares of the errors of estimate shall be a minimum. In short, we desire values

for A, B_1, and B_2 in the equation

$$X_1' = A + B_2X_2 + B_3X_3$$

or, equivalently, the values b_2 and b_3 in the deviation-unit equation

$$x_1' = b_2x_2 + b_3x_3$$

such that the sum

$$\Sigma (X_1-X_1')^2 = \Sigma (x_1-x_1')^2$$

is a minimum.

The task of derivation is somewhat simplified if we transform all three sets of values into standard-score form, i.e., if we set $Z_i = (X_i-M_i)/S_i$. Then our equation becomes

$$Z_1' = B_2^* Z_2 + B_3^* Z_3 \qquad (A2.1)$$

where B_i^* represents <u>the partial regression coefficient in standard-score form</u>. As for the bivariate case, these regression weights are usually called beta coefficients or standardized regression coefficients and denoted by the Greek letter, Beta. Since we are changing the size of our unit of measure, it should be noted that, say B_2^* will not necessarily equal $B_2 = b_2$. Now we need to determine the value of B_2^* and B_3^* such that the average of the squared errors, or

$$\frac{1}{N} \Sigma (Z_1 - Z_1')^2$$

shall be a minimum. Since $Z_1 - Z_1' = Z_1 - B_2^* Z_2 - B_3^* Z_3$, the function to be minimized is

$$f = \frac{1}{N} \Sigma (Z_1 - B_2^* Z_2 - B_3^* Z_3)^2$$

As in the bivariate case, the calculus is used to determine the values of B_2^* and B_3^* which will make this function a minimum. Taking the partial derivative of the function first with respect to B_2^* and then with respect to B_3^*

$$\frac{d_p f}{d_p B_2^*} = \frac{-2 \Sigma Z_2}{N} (Z_1 - B_2^* Z_2 - B_3^* Z_3)$$

$$\frac{d_p f}{d_p B_3^*} = \frac{-2 \Sigma Z_3}{N} (Z_1 - B_2^* Z_2 - B_3^* Z_3)$$

These two derivatives must be set equal to zero and then solved simultaneously for the unknowns, B_2^* and B_3^*. By performing the indicated multiplications, suming, and dividing each equation by 2, we get

$$\frac{-\Sigma Z_1 Z_2}{N} + B_2^* \frac{\Sigma Z_2^2}{N} + B_3^* \frac{\Sigma Z_2 Z_3}{N} = 0$$

$$\frac{-\Sigma Z_1 Z_3}{N} + B_2^* \frac{\Sigma Z_2 Z_3}{N} + B_3^* \frac{\Sigma Z_3^2}{N} = 0$$

Noting that the sum of squares of standard-scores is unity, whereas any sum of cross products of standard-scores divided by N is the correlation between the two variables involved in the cross products, we have by application to the above equations

$$-r_{12} + B_2^* + B_3^* r_{23} = 0$$

$$-r_{13} + B_2^* r_{23} + B_3^* = 0$$

or

$$B_2^* + r_{23}B_3^* - r_{12} = 0$$
$$r_{23}B_2^* + B_3^* - r_{13} = 0$$

Since the rs in the above equations are determinable for any specific set of data, we may treat them as knowns leaving only the B*s as unknowns. Solution of these two simultaneous equations in two unknowns gives

$$B_2^* = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$
$$B_3^* = \frac{r_{13} - r_{13}r_{23}}{1 - r_{23}^2} \qquad \text{(A2.2)}$$

As soon as the values of B_2^* and B_3^* have been determined, they can be substituted in the prediction equation

$$Z_1' = B_2^* Z_2 + B_3^* Z_3$$

so that for a given pair of Z_2 and Z_3 values we can predict the standard-score on the criterion variable. However, it is often more convenient to deal with deviation- or raw-scores. Hence, by replacing the Zs in the preceding equations by their values in terms of raw-scores, means, and standard deviations, we will have

$$\frac{X_1' - M_1}{S_1} = B_2^* \frac{X_2 - M_2}{S_2} + B_3^* \frac{X_3 - M_3}{S_3}$$

or

$$\frac{X_1'}{S_1} - \frac{M_1}{S_1} = B_2^* \frac{X_2}{S_2} - B_2^* \frac{M_2}{S_2} + B_3^* \frac{X_3}{S_3} - B_3^* \frac{M_3}{S_3}$$

After multiplying by S_1 and rearranging terms, we have

$$X_1' = B_2^* \frac{S_1}{S_2} X_2 + B_3^* \frac{S_1}{S_3} X_3 + \left(M_1 - B_2^* \frac{S_1}{S_2} M_2 - B_3^* \frac{S_1}{S_3} M_3\right) \quad \text{(A2.3)}$$

as the regression equation in raw-score form. From this equation, we see that our original B_2 must equal $B_2^* (S_1/S_2)$, $B_3 = B_3^* (S_1/S_3)$, and A = the parentheses term. We may also derive values for the partial regression weights in terms of bivariate regression weights from these equations as follows

$$B_2 = \frac{S_1}{S_2} B_2^*$$

$$= \frac{S_1}{S_2} \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}$$

and, since $r_{23} = B_{23}B_{32}$ by formula (A1.7)

$$= \frac{S_1}{S_2} \frac{r_{12} - r_{13}r_{23}}{1 - B_{23}B_{32}}$$

$$= \frac{\frac{S_1}{S_2} \frac{\Sigma x_1x_2/N}{S_1S_2} - \frac{S_1}{S_2} \left(\frac{\Sigma x_1x_3/N}{S_1S_3} \cdot \frac{\Sigma x_2x_3/N}{S_2S_3}\right)}{1 - B_{23} B_{32}}$$

$$= \frac{\frac{\Sigma x_1x_2/N}{S_2^2} - \frac{\Sigma x_1x_3/N}{S_3^2} \cdot \frac{\Sigma x_2x_3/N}{S_2^2}}{1 - B_{23}B_{32}}$$

or, since $B_{12} = \frac{\Sigma x_1x_2/N}{S_2^2}$ by formula (A1.6), this reduces to

$$B_2 = \frac{B_{12} - (B_{13})(B_{32})}{1 - B_{23} B_{32}} \quad \text{(A2.4)}$$

Thus, we have an equation for the raw- or deviation-score partial regression coefficients in terms of bivariate raw- or deviation-score regression coefficients. The notation for this coefficient is often written $B_{12 \cdot 3}$ to indicate that it is the regression of variable 1 on variable 2, controlling for variable 3 (the expected change in variable 1, given a unit of change in variable 2, controlling for variable 3). Generalizing this notation, we may write

$$b_{ij \cdot k} = B_{ij \cdot k} = \frac{B_{ij} - (B_{ik})(B_{kj})}{1 - (B_{jk})(B_{kj})} \qquad \text{(A2.5)}$$

as the equation for the partial regression coefficient in raw- or deviation-score form of variable i on variable j, controlling for variable k.

We can ascertain the accuracy of the prediction of X_1 from the best combination of X_2 and X_3 by examining the error of prediction, i.e., $X_1 - X_1'$ or $S_1(Z_1 - Z_1')$. The sum of the squares of the errors divided by N will yield the variance of the errors. The square root of this variance would correspond to the standard error of estimate. Let $S_{z1.23}$ be the standard error (in Z-units) for predicting X_1 from X_2 and X_3, i.e., let $S_{z1.23}$ be the standard deviation of the residual terms (in z-units). Then

$$S^2_{z1.23} = \frac{\Sigma (Z_1 - Z_1')^2}{N}$$

$$= \frac{\Sigma (Z_1 - B^*_2 Z_2 - B^*_3 Z_3)^2}{N}$$

$$= \frac{\Sigma Z_1^2}{N} + B_2^{*2}\frac{\Sigma Z_2^2}{N} + B_3^{*2}\frac{\Sigma Z_3^2}{N} - \frac{2B_2^*\Sigma Z_1Z_2}{N}$$

$$- \frac{2B_3^*\Sigma Z_1Z_3}{N} + \frac{2B_2^*B_3^*\Sigma Z_2Z_3}{N}$$

$$= 1 + B_2^{*2} + B_3^{*2} - 2B_2^* r_{12} - 2B_3^* r_{13} + 2B_2^* B_3^* r_{23} \qquad \text{(A2.6)}$$

which by algebraic manupulation reduces to

$$S^2_{z1.23} = 1 - (B_2^* r_{12} + B_3^* r_{13}) \qquad \text{(A2.7)}$$

in terms of standard scores. Of course, S_1^2 times this would give the error variance for raw-scores.

We proceed to define the <u>multiple correlation coefficient</u> as the correlation between Z_1 and the best estimate of Z_1 from a knowledge of Z_2 and Z_3. In symbols,

$$R_{1.23} = R_{z_1 z_1'} = \frac{\Sigma Z_1 Z_1'}{N S_{z_1} S_{z_1'}}$$

$$= \frac{\Sigma Z_1 (B_2^* Z_2 + B_3^* Z_3)}{N S_{z_1'}} \qquad \text{(A2.8)}$$

Note that, for a sample of values $S_{z_1} = 1$. However, it does not follow that $S_{z_1'} = 1$. In order to evaluate this last S, we may think of Z_1 as being made up of two parts - that which we can estimate plus a residual:

$$Z_1 = Z_1' + Z_{1.23}$$

It can be shown that these two parts are independent of each other. Hence, their variances are additive:

$$S^2_{z1} = S^2_{z1'} + S^2_{z1.23}$$

or

$$1 = S^2_{z1} + S^2_{z1.23}$$

then

$$S^2_{z1} = 1 - S^2_{z1.23}$$

However, $S^2_{z1.23}$ is nothing more than the variance of the prediction errors as given by (A2.7); hence this becomes by substitution of (A2.7)

$$S_{z1}' = \sqrt{B_2^* r_{12} + B_3^* r_{13}} \qquad \text{(A2.9)}$$

Then, by substituting (A2.9) in (A2.8), we have

$$R_{1.23} = \frac{\Sigma Z_1 (B_2^* Z_2 + B_3^* Z_3)}{N \sqrt{B_2^* r_{12} + B_3^* r_{13}}}$$

$$= \frac{B_2^* \Sigma Z_1 Z_2 + B_3^* \Sigma Z_1 Z_3}{N \sqrt{B_2^* r_{12} + B_3^* r_{13}}}$$

$$= \sqrt{B_2^* r_{12} + B_3^* r_{13}} \qquad \text{(A2.10)}$$

From formula (A2.10) and (A2.7), we see that we can write the standard error of estimate in raw-score form as

$$S_{1.23} = S_1 \sqrt{1 - R^2_{1.23}} \qquad \text{(A2.11)}$$

This formula may be used to define the multiple correlation coefficient. The relationship is

$$R^2_{1.23} = 1 - \frac{S^2_{1.23}}{S^2_1} = 1 - \frac{S^2_{1.23}}{S^2_1} \qquad \text{(A2.12)}$$

By substituting from (A2.7), we again have (A2.10).

At this point, we may note the similarity of formula (A2.11) to the standard error of estimate for the bivariate situation. Thus, the interpretation of the correlation coefficient in terms of reduction in the error of estimate holds for the multiple correlation coefficient in exactly the same manner as for the ordinarry bivariate correlation coefficient. Furthermore, the interpretation in terms of proportion of variance explained also holds for the multiple correlation coefficient. However, in the case of two predictor variables, we find some interesting differences. Let us explore those peculiarities.

We must answer the question as to the relative importance of the two predictor variables as contributors to variation in the criterion variable. Obviously, the B coefficients in the raw-score regression equation cannot be interpreted as indicating the relative contribution of the two independent variables since the two B coefficients usually involve different units of measurement. Therefore, a B_2 twice as large as B_3 does not imply that B_2 is twice as important as B_3. However, the variables in standard-score form will be comparable and hence the beta coefficients in the standard-score form of the regression equation will be comparable. Since

$$s^2_{z_1} = s^2_{\hat{z}_1} + s^2_{1.23}$$

or

$$1 = s^2_{\hat{z}_1} + s^2_{z1.23}$$

and

$$1 - s^2_{z1.23} = R^2_{1.23}$$

it follows that

$$R^2_{1.23} = S^2_{z'_1}$$

In words, $R^2_{1.23}$ which corresponds to the proportion of variance explained by variables 2 and 3 is equal to $S^2_{z'_1}$, or the variance of the predicted standard-scores.

On the other hand, note that, since

$$z'_1 = B^*_2 z_2 + B^*_3 z_3$$

we can indicate the value of $S^2_{z'_1}$ as

$$R^2_{1.23} = S^2_{z'_1} = \frac{\Sigma (z'_1)^2}{N} = \frac{\Sigma (B^*_2 z_2 + B^*_3 z_3)^2}{N}$$

$$= \frac{B^{*2}_2 \Sigma z^2_2 + B^{*2}_3 \Sigma z^2_3 + 2B^*_2 B^*_3 \Sigma z_2 z_3}{N}$$

which is

$$R^2_{1.23} = S^2_{z'_1} = B^{*2}_2 + B^{*2}_3 + 2B^*_2 B^*_3 r_{23} \qquad \text{(A2.13)}$$

In short, the predicted variance, which corresponds to the "explained" variance, can be broken down into three additive components. Furthermore, we see that the relative importance of the variables X_2 and X_3 in "explaining" variation in X_1 can be judged by the magnitude of the squares of the beta coefficients. The third term in formula (A2.13) represents a joint contribution which is a function of the amount of correlation between the two predicting variables.

A.2.2. More Than Three-Variables. The extension of multiple correlation and regression to include any number of variables involves

the same principles as for the three-variable case. In other words, the interpretation of the regression and correlation coefficients is the same for n as for 3 variables, and the extension of the formulas should be obvious.

BIBLIOGRAPHY

Blalock, Hubert M. Causal Inferences in Nonexperimental Research. Chapel Hill: University of North Carolina Press, 1964.

Duncan, Otis Dudley. Path Analysis: Sociological Examples. American Journal of Sociology, Vol. 72, July 1966, pp. 1 - 16.

Eaton, David C. The Use of Causal Models in the Study of Low Income Status. Unpublished Ph.D. dissertation, The University of Texas, Austin, Texas, 1967.

Kempthorne, Oscar. An Introduction to Genetic Statistics. New York: John Wiley and Sons, 1957, chapter 14.

McNemar, Quinn. Psychological Statistics. New York: John Wiley and Sons, 1962.

Simon, Herbert A. Models of Man. New York: John Wiley and Sons, 1957, chapters 1 and 2.

Tukey, J. W. Causation, Regression and Path Analysis. in O. Kempthorne et. al. (eds.), Statistics and Mathematics in Biology. Ames: Iowa State College Press, 1954, chapter 3.

Turner, Malcolm E. and Charles D. Stevens. The Regression Analysis of Causal Paths. Biometrics, Vol. 15, June 1959, pp. 236-258.

Wold, Herman and Lars Jureen. Demand Analysis. New York: John Wiley and Sons, 1953.

Wright, Sewall. Correlation and Causation. Journal of Agricultural Research. Vol. 20, 1921, pp. 557-585.

Wright, Sewall. The Method of Path Coefficients. Annals of Mathematical Statistics. Vol. 5, September 1934, pp. 161-215.

Wright, Sewall. The Interpretation of Multivariate Systems. in O. Kempthorne et.al. (eds.), Statistics and Mathematics in Biology. Ames: Iowa State College Press, 1954, chapter 2.

Wright, Sewall. Path Coefficients and Path Regressions: Alternative or Complimentary Concepts? Biometrics, Vol. 16, June 1960, pp. 189-202.

Wright, Sewall. The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis. Biometrics, Vol. 16, September 1960, pp. 423-445.

F-OEO
PS

FROM:
ERIC FACILITY
SUITE 601
1735 EYE STREET, N. W.
WASHINGTON, D. C. 20006